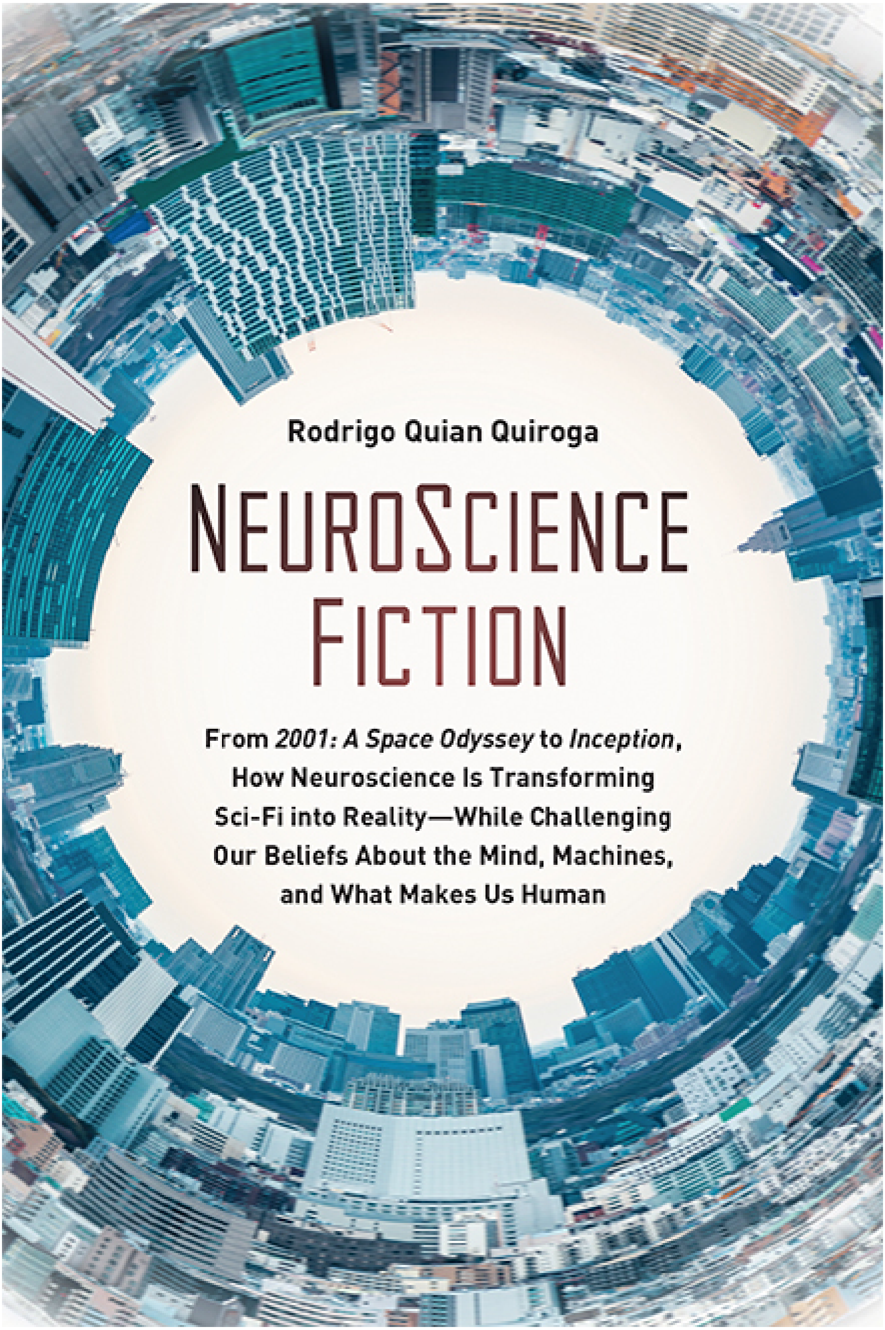


Rodrigo Quian Quiroga

NEUROSCIENCE FICTION

From *2001: A Space Odyssey* to *Inception*, How Neuroscience Is Transforming Sci-Fi into Reality—While Challenging Our Beliefs About the Mind, Machines, and What Makes Us Human

Praise for *NeuroScience Fiction*

"If you like science fiction, this book is for you, but if you like science *and* fiction, then this is certainly your book. Intelligent and well informed."

—Antonio Damasio, director of the Brain and Creativity Institute, University of Southern California, and author of *Descartes' Error* and *The Strange Order of Things*

"This exceptional book, based on neuroscientific facts, beautifully links philosophy, literature, and fiction, and brings to mind the expressive inscription on the frieze of the Alte Oper in Frankfurt, 'Dem Wahren, Schönen, Guten' ('To the true, the beautiful, the good'). Rodrigo Quian Quiroga has written a thorough, provocative answer to the deep philosophical question of what makes us human. His truly remarkable book blends art, science, and philosophy as seen through the highly original lens of brain research and movies."

—Gustavo Deco, ICREA professor, director of the Center for Brain and Cognition, Theoretical and Computational Group, Universitat Pompeu Fabra, Barcelona

"Of all of Mother Nature's magic tricks, the conjuring of conscious sensations, thoughts, and feelings from a small vat of cells is arguably the most wondrous of all. For a long time, Descartes told us the recipe was lost. But now, renowned cognitive neuroscientist Rodrigo Quian Quiroga invites us on an exhilarating journey, in which the motley trio of neuroscience, philosophy, and Hollywood together seek after the recipe of the mind. Quian Quiroga, who himself discovered what is sure to be an essential ingredient, the concept cells, shows us how frighteningly close we are getting and how huge the stakes are."

—Doris Y. Tsao, neuroscientist and professor of biology at the California Institute of Technology

"This book is a fascinating journey across some of the most influential science-fiction movies that deal with the deepest questions in neuroscience and philosophy of mind. Only a Renaissance man like Rodrigo Quian Quiroga could find these deep connections between fiction and human nature and show how they are becoming reality."

—Jose M. Carmena, professor of electrical engineering and neuroscience, University of California, Berkeley

"Science-inspired fiction hooks so many of us because of what it has to say about our immediate potential—both for greatness and for decline. *NeuroScience Fiction* will be a treat for sci-fi movie fans and anyone who has wondered just how close we are to the horizon technologies—which is often much closer than we realize."

—James H. Fallon, professor at the Department of Psychiatry and Human Behavior, University of California, Irvine, and bestselling author of *The Psychopath Inside*

NeuroScience Fiction

NeuroScience Fiction

From *2001: A Space Odyssey* to *Inception*, How Neuroscience Is Transforming Sci-Fi into Reality—While Challenging Our Beliefs About the Mind, Machines, and What Makes Us Human

Rodrigo Quian Quiroga

BenBella
BenBella Books, Inc.
Dallas, TX

NeuroScience Fiction copyright © 2020 by Rodrigo Quian Quiroga

All rights reserved. No part of this book may be used or reproduced in any manner whatsoever without written permission of the publisher, except in the case of brief quotations embodied in critical articles or reviews.

Originally published in Spanish as *NeuroCienciaFicción: Cómo el cine se adelantó a la ciencia* in 2018 by Penguin Random House Grupo Editorial, S.A. Copyright © 2018 by Rodrigo Quian Quiroga.

Chapter opening illustrations by Candela Insua
Translation by Cecilia Molinari

BenBella Books, Inc.
10440 N. Central Expressway, Suite 800
Dallas, TX 75231
www.benbellabooks.com
Send feedback to feedback@benbellabooks.com

BenBella is a federally registered trademark.

First E-Book Edition: April 2020

Library of Congress Control Number: 2019054170
ISBN 9781950665051 (print)
ISBN 9781950665228 (electronic)

Editing by Laurel Leigh and Alexa Stevenson
Copyediting by Scott Calamar
Proofreading by Jenny Bridges and Lisa Story
Indexing by WordCo Indexing Services, Inc.
Text design and composition by Aaron Edmiston
Cover design by Sarah Avinger
Cover photo © Shutterstock / metamorworks
Printed by Lake Book Manufacturing

Distributed to the trade by Two Rivers Distribution, an Ingram brand www.tworiversdistribution.com

Special discounts for bulk sales are available.
Please contact bulkorders@benbellabooks.com.

For little princess Maia,
and for Rodriguito and Felipe,
her two guardian princes.

CONTENTS

Introduction

1 *2001: A Space Odyssey*—Machine Intelligence
2 *Blade Runner*—Can Androids Feel?
3 *Planet of the Apes*—Animal Consciousness
4 *The Matrix*—The Illusion of Reality
5 *Until the End of the World*—Reading the Mind
6 *Minority Report*—Free Will
7 *RoboCop*—Cyborgs and Identity
8 *Inception*—Dream Construction
9 *Total Recall*—Memory Manipulation and Implantation
10 *Open Your Eyes (Vanilla Sky)*—Immortality

Epilogue: The Dawn of a New Philosophy
Acknowledgments
Notes
Index
About the Author

INTRODUCTION

Movies go far beyond imagining the obvious. They tend to be ahead of their time, proposing disruptive and fascinating scenarios. In *Inception*, a group of conspirators implant false memories. In *Until the End of the World*, a crazy scientist is able to read dreams. And in *2001: A Space Odyssey*, a supercomputer feels and thinks like a person.

Just as sci-fi movies tend to rely on the latest scientific breakthroughs, science is inspired by the prolific imagination of filmmakers and, years after these blockbusters first appeared in theaters, we find ourselves immersed in those realities. Implanting memories using optogenetics and the resulting possibility of voluntarily manipulating groups of neurons, reading the mind during sleep through advanced decoding algorithms, or getting computers to surpass human thought in infinite tasks through deep neural networks are all achievements that have been described in prestigious scientific journals in recent years.

That's what this book is about: how science is achieving what decades ago seemed impossible, and how these advances are making us reconsider the big questions that we humans have been asking ourselves for as long as we can remember. In *NeuroScience Fiction*, I want to share with you what keeps me awake at night, what fascinates me. Our world is experiencing a revolution in neuroscience—so profound, so close to the roots of our own essence, that it feeds off the most transcendental philosophical discussions and, at the same time, so captivating that it has fired up the imagination of futurist writers as an endless source of sci-fi films.

This book is supported by that tripod—three passions that reinforce each other: science fiction, neuroscience, and philosophy. Movies such as *2001, Blade Runner, The Matrix*, and *Inception*, among many others, make us think about human consciousness and an eventual consciousness of machines, as well as the existence of external reality, free will, and the sense of identity that makes us recognize ourselves as people, to attribute to ourselves the notion of a self, and even to consider if one day we could become immortal.

Neuroscience is advancing at an alarming rate, and many of the ideas in these movies are increasingly related to science and less so to fiction. Although it may be hard to believe, implanting a memory, reading the mind, communicating with patients in a coma, predicting a person's decision just seconds before they make it, creating cyborgs, or getting a paralyzed person to walk again are no longer pipe dreams. These are solid and verifiable realities that have stricken us neuroscientists with awe.

I am fortunate enough to have been working in neuroscience for years now and also had the luck of discovering a type of neuron called *concept cells*, or "Jennifer Aniston neurons," which respond to specific concepts, such as the aforementioned actress, the character Luke Skywalker, or the soccer player Diego Maradona. These neurons give us concrete evidence about how our memories and thoughts work, and I also believe they have staggering connotations.

Almost four hundred years ago, the French philosopher René Descartes proposed a division between the body and the mind, arguing that communication between both happens through the pineal gland, which was believed to exist only in humans. Nowadays, the idea of conceiving the mind as a separate entity from the brain has lost support among scientists; what's more, we also know that other animals have a pineal gland.

So, what makes us human? What differentiates us and makes us the dominant species? Could monkeys really subdue humans like in *Planet of the Apes*? Is there something special in our brain that makes us more intelligent and leads us to write books and wonder about these things?

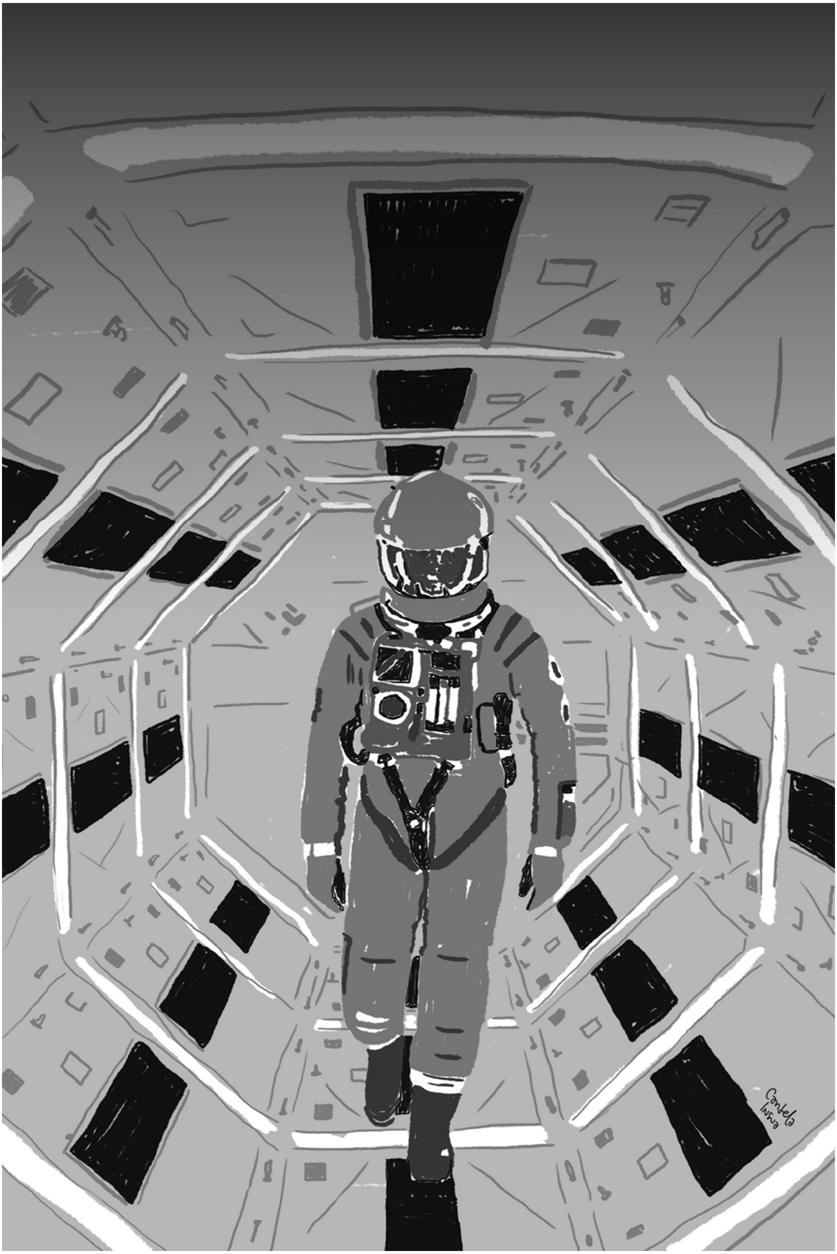
Going back to one of the big questions that has led to a centuries-long debate in philosophy, I dare to postulate that the concept cells I identified a few years back might be the neural substrate that gives way to our most elaborate and abstract thoughts, those that make us human and differentiate us from other animals and from computers. In the context of intersections between science and science fiction in this book, and giving my imagination free rein, I even dare to speculate that inducing such neurons and the processes they may give rise to could eventually make some of the most fascinating sci-fi predictions a reality.

—Rodrigo Quian Quiroga

CHAPTER 1

2001: A SPACE ODYSSEY

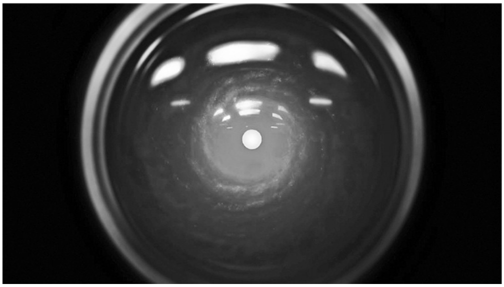
Machine Intelligence


Candela
Insua

In April 1968, a year before a human stepped on the moon, Stanley Kubrick premiered his monumental *2001: A Space Odyssey*, which he wrote with Arthur C. Clarke, renowned sci-fi author, who would subsequently publish a novel with the same title.

Half a century after its release, *2001* continues to be considered by many as the most accomplished sci-fi movie. It opens with three minutes of absolute darkness, establishing a point of reference from where Kubrick develops a unique audiovisual experience. Most of the movie is dominated by majestic scenes in space with unprecedented realism for the time. Breaking from a stereotypical story line, the scenes are accompanied by classical music and are nearly absent of dialogue, with endlessly sustained shots that generate an extremely slow cadence, evoking the rhythm of gravity-less events. The first word is heard close to thirty minutes into the movie, and most of the plot takes place in silence, without innocuous conversations ruining the incredible experience proposed by Kubrick.

Beyond the realism and beauty of these scenes, the movie transcends for another reason. Divided into three parts, the second act describes the monotonous routine of two astronauts (Dave Bowman and Frank Poole) in charge of a space mission to Jupiter, and shows us another three astronauts in hibernation. Although, by a long shot, the character who steals the show is the sixth crew member, HAL 9000, a supercomputer represented by a calm, monotonic voice and a single red lens.



HAL 9000, the supercomputer in *2001: A Space Odyssey*.

Kubrick's outstanding achievement is to convince his audience that the red light is a sentient being, with emotions, fears, and interests. As in any good story line, the characters evolve as the plot unfolds, and the once perfect companion on this mission to Jupiter, seemingly incapable of the slightest mistake, morphs into a frightening being that makes erratic, unpredictable choices due to something as human as the fear of ceasing to exist.

Afraid of being shut down by Bowman—who has ventured outside the main spacecraft in a small, one-man vehicle known as the EVA pod and now wants to come back in—HAL makes an unexpected, though logical, decision, considering that it prioritizes its survival instinct:

> **Bowman:** Open the pod bay doors, HAL.
> **HAL:** I'm sorry, Dave. I'm afraid I can't do that.
> **Bowman:** What's the problem?
> **HAL:** I think you know what the problem is just as well as I do.
> **Bowman:** What are you talking about, HAL?
> **HAL:** This mission is too important for me to allow you to jeopardize it.
> **Bowman:** I don't know what you're talking about, HAL.
> **HAL:** I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.
>
> —*2001: A Space Odyssey*

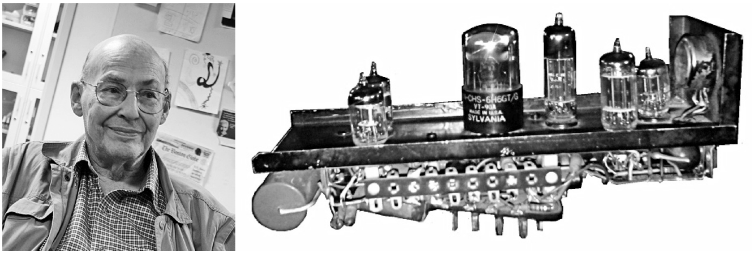
With a terrifying coldness, HAL does not hesitate to kill whoever threatens its existence. Its arguments are rational and irrefutable—after all, it is a computer—yet when the disconnection is inevitable, HAL shows its sensitive and human side:

> **HAL:** Stop, Dave. I'm afraid. I'm afraid, Dave. Dave, my mind is going. I can feel it. I can feel it. My mind is going. There is no question about it. I can feel it. I can feel it. I can feel it. I am a . . . fraid.
>
> —*2001: A Space Odyssey*

The fact that a computer—ultimately, a collection of algorithms—can think, exercise its own will, and even feel fear or refuse to accept orders, opened up an intriguing and surprising discussion at the time, considering that in the 1960s we were barely witnessing the dawn of digital calculators, while the first personal computers were introduced in the mid-1970s. However, Kubrick and Clarke's genius glimpse into the future of computers had an extraordinary scientific sustenance. Faced with danger, HAL decides to sabotage the life support of the hibernating astronauts, one of them being Victor Kaminski, whose name is a play on words in honor of the late Marvin Minsky, the great pioneer of artificial intelligence (AI), who was an adviser for the movie.

Among many other achievements, Minsky is recognized for creating one of the first artificial neural networks, known as *SNARC* (Stochastic Neural Analog Reinforcement Computer), in 1951. This network of forty neurons was basically composed of vacuum valves, circuits, and motors, and it simulated the behavior of a mouse searching for the exit in a maze. Later, Minsky would establish the famous Artificial Intelligence Lab at MIT (Massachusetts Institute of Technology), and a while later, in 1986, he published *The Society of Mind*, where he proposed a radical view on intelligence and the eventual possibility of thinking computers.[1]

Minsky considered intelligence to be the ability to solve difficult problems, which, in turn, can be segmented into a collection of simpler mechanical processes called *agents*, which do not require thinking or consciousness. For example, according to Minsky, the process of drinking a cup of tea involves agents that take care of grabbing the cup so that it doesn't fall; others that keep it balanced so as not to spill the tea; agents that transmit a thirsty feeling, causing the person to want to drink; and agents involved in motor movements to bring the cup to the mouth.[2]



Marvin Minsky and the circuit of one of the forty neurons in the neural network SNARC.[*]

Considering that a machine can carry out these simple and unconscious processes, Minsky argued that there is, in principle, no fundamental difference between human intelligence and the intelligence a computer could have. In other words, "there is no clear boundary between psychology and Artificial Intelligence because the brain itself is a kind of machine."[3]

The barrier that separates artificial and human intelligence is becoming increasingly vague; however, before reviewing AI's achievements, especially in recent years, let's further consider this last affirmation, as it involves a philosophical discussion that has been going on for more than two thousand years.

The big question brought up by HAL 9000 and Minsky's view is whether or not a machine could manage to think and be intelligent. A priori, the answer seems to be a decisive no. A machine can execute complex algorithms and carry out several tasks—such as performing a calculation—much better than a human, but the machine responds to instructions given by a human user and, in principle, cannot think for itself because it does not possess a mind. But what is the mind? What is it made of? What makes it an exclusively human attribute?

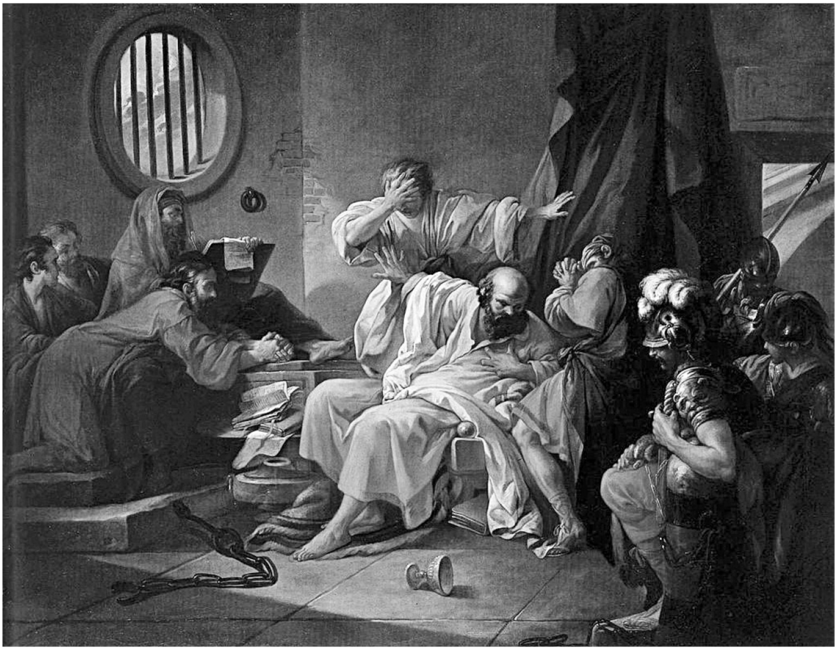
More than 2,400 years ago, in 399 BC, an Athenian court declared Socrates guilty of corrupting young Athenians and sentenced him to death. The unjust penalty was imposed because Socrates questioned the existing order, doubting and reconsidering absolutely everything, and because he was openly defiant toward his prosecution process.

Socrates's thoughts during his last days before drinking from the infamous cup of hemlock come to us from two dialogues written by Plato (427–347 BC), his most notable student. In *Crito*, Plato writes that Socrates refused to be saved by his friends, arguing with astonishing rationale, almost as if he were talking about someone other than himself, that he preferred death to escaping Athens and betraying his principles. The other dialogue that refers to Socrates's death is *Phaedo*, a treatise on the soul in which Plato describes Socrates's last moments and the stoicism with which he accepts his fate, contrary to his followers' distress.

The coolheadedness with which Socrates faces death derives from his unwavering intellectual coherence and his firm conviction of the existence of an immortal soul. Plato attributes to his teacher the notion of the existence of universal concepts—"equality," "beauty," or "kindness"—that are within the purview of the soul and that can be understood only through thought. Literal things, such as "men," "horses," "suits," or "furniture," correspond to the mutable domain of the material that can be perceived with the senses. Plato says:

> When the soul uses the body to consider an object, either through sight or hearing or

> any other sense (because the body's only function is to tend to objects through the senses), she then finds herself attracted through the body to things that are never the same; she gets lost, confused, hesitates, and becomes dizzy, as if she were intoxicated . . . But when she examines things by herself, without resorting to the body, she heads to what is pure, eternal, immortal, unchangeable; and since she is of the same nature, she is united and embraces with it all she can and gives of herself her own nature. Then she no longer goes astray, she remains the same, because she is united to that which never changes, and participates in its nature; and this state of the soul is called wisdom.
>
> —Plato, *Phaedo* (53, 55)[4]



The Death of Socrates (oil on canvas by Jacques-Philip-Joseph de Saint-Quentin). In the center of the scene, Socrates, after drinking from the hemlock cup, rejects his disciples' lament.

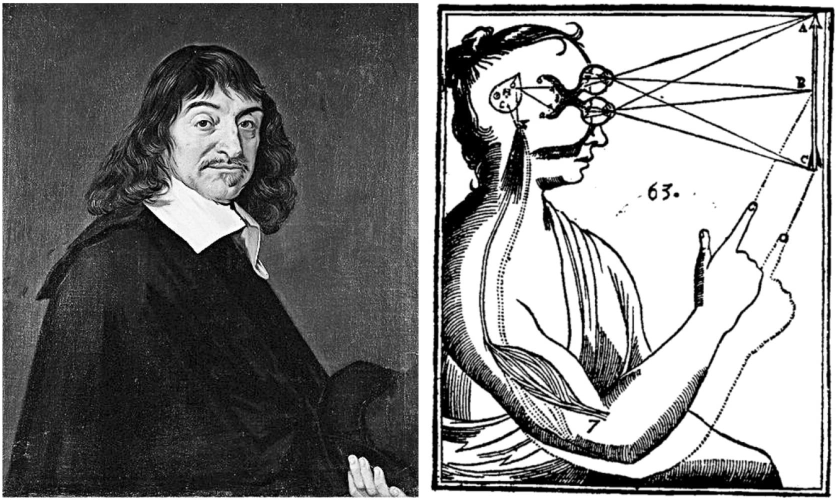
Plato views the mind as a separate entity from the body, and the philosopher's job is then to separate himself from the confusions imposed by the perception of physical reality in order to achieve essential knowledge. The philosopher maintains a pure soul, and in death he frees himself from the body's anchor and the temptations and deceptions of the senses, to move on to another world in the company of gods. Hence Socrates's disappointment upon seeing his disciples clinging to the notion of saving his life and failing to understand his most fundamental lesson. But what does this philosophical argument have to do with Minsky's ideas about AI? If Plato's contentions are true, built from ideology for which Socrates did not hesitate to give up his life, then there is a fundamental difference between humans and machines. Humans have a mind, a soul, while machines are only matter.

René Descartes (1596–1650), known as the father of modern philosophy, picked up Plato’s ideas and clearly differentiated mind from matter, laying the groundwork for what is known as *Cartesian dualism*.

> And although perhaps (or better yet, with absolute certainty, as I will immediately say) I have a body I am closely connected to, with everything, given that, on one hand, I have a clear and distinct idea of myself, in so far as I am just a thing that thinks—and not extensive—and, on the other hand, I have a distinct idea of the body, in terms of it being only an extensive thing—and not thinking—it is true, then, that that self (in other words, my soul, the reason I am who I am), is entirely separate from my body, and can exist without it.
>
> —René Descartes, *Meditations on First Philosophy* (Sixth Meditation)[5]

For Descartes, the brain deals with mechanical acts in both humans and animals, while the exclusively human soul nurtures mental processes. The body is a sophisticated machine that acts in the material domain (*res extensa*), while the soul is an intangible entity in charge of thought (*res cogitans*). As mentioned, according to Descartes, the interaction between mind and body happens via the pineal gland, an organ that at the time was erroneously believed to exist only in humans. And therein lies Cartesian dualism’s greatest fault. Not in the fact that the pineal gland doesn’t have the exclusive function that Descartes postulated, but in that Cartesian dualism does not explain how the mind can interact with the brain.

Electromagnetism describes when variations in an electric field generate a magnetic field and, at the same time, the variations in the *magnetic* field generate an *electric* field. Electricity and magnetism are intimately connected, as we know from Maxwell’s equations, so named for the nineteenth-century Scottish physicist James Clerk Maxwell. But is there a similar interaction between mind and matter? At first, it is possible to conceive that the activity of neurons produces mental processes, but how can nonphysical mental processes produce brain activity? How could the mental idea of wanting to move an arm, as in the drawing on page 14, affect the firing of neurons to carry out the movement? Cartesian dualism does not have an answer to this question.*

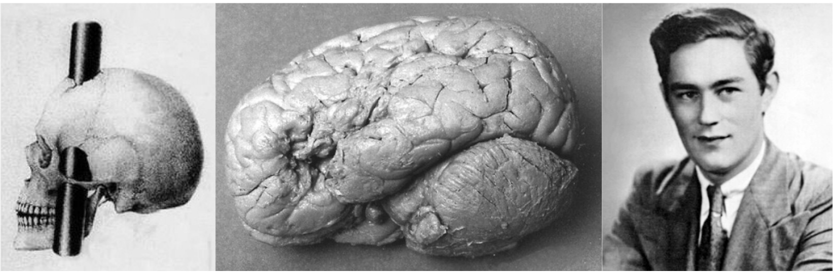


René Descartes (oil on canvas by Frans Hals, Louvre Museum) and one of his sketches illustrating the pineal gland's function. Light travels through the eyes to the pineal gland, and there the physical stimuli become thoughts that lead to intelligent actions, like moving an arm toward a specific point.

These days, we know that the mind is not a separate entity from the brain. In fact, irrefutable scientific evidence shows how certain brain injuries can change a person's "mind." One of the most renowned cases is that of Phineas Gage, a railroad construction foreman, who in 1848 suffered an accident in which an explosion fired an iron rod straight through his skull, destroying his brain's left frontal lobe. The surprising fact was not only that Gage didn't die, but that the brain injury produced a dramatic change in his personality. He went from being a responsible and friendly worker to behaving in an irreverent, salacious, and aggressive manner, as if he had lost the ability to censor himself.[6]

Another well-known case in neuroscience is that of Louis Victor Leborgne, a Parisian cobbler whom hospital staff nicknamed "Monsieur Tan," since that was the only word he could say. In 1861, Leborgne would die one week after being examined by Pierre Paul Broca, a brilliant surgeon and anatomist who, in a postmortem autopsy, discovered a noticeable lesion in the left frontal lobe. This is what we now call Broca's area (the first irrefutable scientific evidence about the localized functions in the brain), which is known to be fundamental for language.[*] A third case worth mentioning is that of Henry Molaison (also known by his initials H.M.), perhaps the most studied patient in the history of science.

H.M. suffered epileptic seizures that could not be treated with medication. In 1953, neurosurgeon William Scoville surgically removed his hippocampus (a seahorse-shaped structure that is usually compromised by epileptic seizures) and adjacent areas in both cerebral hemispheres. The unexpected result was that, after the surgery, H.M. was unable to recognize the hospital staff, find his way to the bathroom, or recall daily events. In other words, H.M. lost his ability to form new memories, which established the importance of the hippocampus for memory.[7]



Reconstruction of the lesion suffered by Phineas Gage (*left*), Leborgne’s postmortem brain (*center*), and photo of Henry Molaison, who had both hippocampi removed to try to cure his epilepsy.

Summing up the previous discussion, on one hand, Cartesian dualism cannot explain how the mind interacts with matter, and, on the other hand, functions that at first seemed exclusive to the mind, like language, memory skills, and even personality traits, are not an independent entity but the activity of the brain.

The coup de grâce to Cartesian dualism would be delivered by philosopher Gilbert Ryle in the mid-twentieth century.[8] Ryle effectively stigmatized Descartes’s notion of the mind as the *ghost in the machine* dogma, based on a classification error. Ryle gave the example of a foreigner visiting the University of Oxford who, after walking across different colleges and departments, wonders, *Where’s the university?* as if the university were the same type of entity as its components. The university, of course, is the organization that groups the different colleges and departments, just like what we call the mind is a group of cerebral processes (and note how this idea corresponds with Minsky’s view on intelligence).

Nowadays, very few scientists still uphold Cartesian dualism. However, the separation between body and mind is deeply rooted in popular thought when we speak about mental illnesses (such as depression, bipolar disorder, or schizophrenia) or, for instance, when we attribute to the mind our will and ability to make decisions to choose our future (we will get into this further in chapter six, when we talk about *Minority Report*). Francis Crick, one of the great scientists of the twentieth century,[*] masterfully summarizes the view the majority of neuroscientists have about the mind and the brain:

> “You,” your joys and your sorrows, your memories and ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules . . .
>
> —Francis Crick, *The Astonishing Hypothesis*[9]

I’ve wandered quite far from Kubrick and his monumental *2001*. Based on a statement by Marvin Minsky about the possibility of artificial intelligence, we took a leap more than two thousand years back to ponder the ideas of Socrates and Plato and, closer in time, of Descartes, Ryle, and contemporary scientists. I have taken a long detour to justify Crick’s assertion. But this viewpoint, which is known as *materialism*, leads to surprising conclusions. Specifically, if the mind is nothing more than the activity of neurons, then we could replace neurons with transistors that, when organized in complex circuits, would replicate brain functions. We could also use algorithms to simulate the behavior of neurons and circuits that are a part of neural networks. The key is in the fact that, having rejected Cartesian dualism, there seems to be nothing that would prevent a machine from being able to become an intelligent being like a human. After all, whether neural activity is based on circuits of carbon (organic matter) or silicon (the inert material used to make computer chips) shouldn't change anything. It is just a matter of technology, though far from trivial, given that the human brain is made up of almost 100 billion neurons and approximately 100 trillion connections among them.

But how far has AI come? In *2001*, HAL 9000 played chess with astronaut Frank Poole, admired drawings by Dave Bowman, controlled the spacecraft's operation, and talked with both crew members. When the movie was released in the 1960s, this was no more than the fruit of Kubrick, Clarke, and Minsky's prolific imagination. But are HAL's accomplishments nowadays real or still fiction?

The most famous advance in artificial intelligence happened at the end of the last century. To the surprise of all humanity, the person who was considered one of its most brilliant exponents, Garry Kasparov, the unbeatable chess champion, was defeated in a series of six games by the supercomputer Deep Blue.

Supported by the extraordinary advances of the last decades, the IBM team that designed Deep Blue got it to calculate around 200 million moves per second and anticipate about twenty moves to choose the most favorable one. But it wasn't all brute force and processing speed. Deep Blue also had an incredibly extensive database with different openings and the development of countless games. Such a combination beat Kasparov and unleashed an animated discussion about whether there is a limit to the advance of computers, and if these could eventually surpass human intelligence.

AI advances are not limited to beating us at chess, exploiting a greater ability for calculations and memory storage. Among other things, AI is used to read handwriting, with a performance almost rivaling that of a human—for example, to automatically classify the addresses of millions of letters in the mail or to process the information on a check inserted into an ATM. Similar programs are also used to recognize (and reproduce) speech. Most of us are indeed familiar with tablets and smartphones, which have built-in applications allowing them to verbally interact with the user, not just by recognizing questions or commands but also processing content and acting accordingly.

In *Borges and Memory*,[10] I stated that a computer can reliably store information but without understanding its content. As an example, I argued that a high-definition TV offers a very precise reproduction of an image, but it cannot interpret if what is playing on the screen is a soap opera, an action movie, a landscape, or a documentary. Nowadays, a few years later, I'm no longer so sure about this.

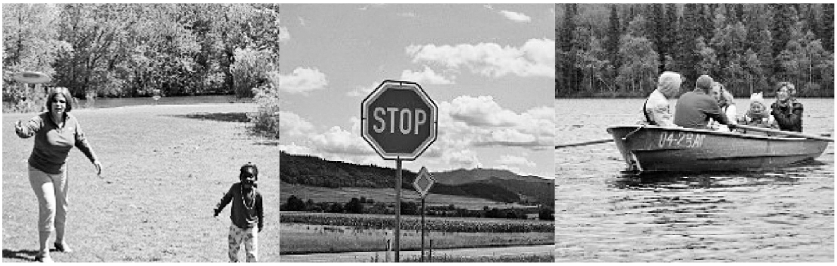
Beyond the fame that came with Deep Blue's defeat of Kasparov, the great AI revolution, a true paradigm shift, happened recently—in 2012. In an automated image classification contest, a group led by Canadian scientist Geoffrey Hinton broke all the records. The contest consisted of classifying around one million internet images in one thousand different categories: "container ship," "leopard," "scooter," "mushroom," etc. The algorithm designed by Hinton and his students had such an advantage that the team that landed second place did so with almost double the errors.[11] Even more striking is the fact that after subsequent improvements, the algorithm used by Hinton's team managed to reach a performance similar to that of a human.[12]

The idea behind Hinton's algorithm is not new. In fact, it was proposed in the 1980s. It is based on using neural networks with several processing layers, where successive layers analyze increasingly complex parameters. Replicating the operation of the visual system, the first layers extract basic characteristics from the image, such as local contrasts; the results are analyzed by higher levels, and so on, hence achieving an increasingly abstract representation of the image.[13] Given the relatively high number of layers used by these networks, they are known as *deep neural networks*.

The network used by Hinton had seven layers (in addition to the first and last ones for the entry and exit of data) and was made up of 650,000 neurons, with around 60 million adjustable parameters. Without needing to be an expert in neural networks, it is clear that the adjustment of so many parameters requires many previous calculations. In other words, to get a deep network to recognize images, it must be fed countless examples. This was exactly the limitation deep networks faced in the 1980s. Back then, they weren't useful in practice because it took too long to train them. However, this limitation recently evaporated after the design of new processors and the use of graphic cards that significantly accelerated the calculation ability. Having resolved the technological issue, it was then possible to train deep networks with millions of examples, and the results demolished the performances obtained through other methods.

Deep networks have become the new prima donna of AI. And these advances have not gone unseen by Silicon Valley giants. Hinton and two of his students were recruited by Google, and other deep network experts now work for Facebook, Microsoft, and IBM. The possible applications defy the limits of imagination; from algorithms for speech or handwriting recognition to the automated classification of images—for example, to tag billions of photos uploaded to the internet—or the development of robots and self-driving automobiles, with systems capable of recognizing and reacting to different situations on the street. It is actually no coincidence that one of the most advanced designs of a self-driving car is made by Google.

In the last few years, one spectacular result has been the possibility of not only recognizing but also interpreting the content of images. An algorithm based on a deep network, connected to another network in charge of speech processing, was able to extract a sense of what different internet photos portrayed. For example, the algorithm wasn't limited to confirming that a photo had an image of a woman and a Frisbee; it also interpreted that the photo showed a woman throwing a Frisbee in the park. Similarly, the algorithm interpreted that a stop sign was on the road and not on the mountain in the backdrop (even though that was literally seen in the photo), and in a third example it recognized that a group of people were sitting in a boat.[14]



Images recognized by algorithms using deep networks.

The most renowned achievement with deep networks happened in late 2015, when a group of Google researchers in London developed AlphaGo (a program with which to play the board game Go) and faced the European Go champion in a five-game match. Reliving the famous Deep Blue and Kasparov duel of two decades earlier, once again a human was defeated by the advances of AI, but, this time, in a much more complex task.

Go is seen as the game that poses the ultimate challenge for AI. The big difference with chess is that while in chess there are around thirty-five possible moves per move, in Go the number of possible moves is approximately 250. So, the number of possible combinations a computer would have to review to explore future positions in a game of Go is immensely higher than the number it needs to review in a game of chess, which makes it literally impossible to use brute force based on calculation speed.[15] The results of the game were revealed in 2016 and had a great impact in the media: AlphaGo had destroyed the European champion five to nil.[16]

A millenary game from ancient China, Go isn't that well-known in Europe. Many experts argued that the European champion wasn't a truly important challenge. Yet, a few months later, in South Korea, AlphaGo would face Lee Sedol, currently one of the most renowned Go players. Sedol believed he would easily win because there was an enormous difference between his ability to play and that of the European champion, who wasn't even ranked among the best three hundred players in the world. Nevertheless, the final outcome would result in AlphaGo winning four victories to one against Sedol.

Sedol was truly astonished by AlphaGo's performance, but what he didn't know was that he had faced a much superior version to the one that defeated the European champion. AlphaGo was initially trained using a database of thirty million games and, after this initial training, the team of scientists made AlphaGo play games against itself to continue learning. While a professional plays around one thousand games a year, in the few months between the match in Europe and the one against Sedol, AlphaGo continued learning at an insane rate, tirelessly playing around one million games a day. Such a frantic learning speed brings to mind another classic sci-fi movie: *The Terminator*. In a scene from the second movie in the series, Terminator explains to his human cohort how Skynet, the AI that subdued humans, was created.

> **Terminator:** The Skynet funding bill is passed. The system goes online on August 4, 1997. Human decisions are removed from strategic defense. Skynet begins to learn, at a geometric rate. It becomes self-aware at 2:14 am Eastern time, August 29. In a panic, they try to pull the plug.
> **Sarah Connor:** Skynet fights back.
>
> —*Terminator 2: Judgment Day*

What's then left for us humans? Should we resign ourselves to the possibility of a superior intelligence? Should we start taking measures to avoid being subdued by what we ourselves have created? Anticipating this discussion by several decades, the fantastic sci-fi writer Isaac Asimov envisioned the consequences of advances in robotics and proposed in *I, Robot* his three famous laws, destined to govern the behavior of robots:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict

with the First or Second Law.[17]

In several interviews, Marvin Minsky (who passed away in early 2016, just days before AlphaGo's first triumph was revealed) stated that the achievements with AI in our time—be it identifying images, recognizing speech, or winning at chess—are far from reproducing human intelligence. As an example, Minsky used to say that any kid knew that he or she could pull an object with a rope, but couldn't push it, while no computer has this knowledge.

Human intelligence goes far beyond the logic that can be implemented in a computer. People use an umbrella to avoid getting wet, said Minsky, but they don't care about getting wet in a swimming pool. The suitcase used to transport clothes can also be used as a makeshift step stool to change a light bulb. This is the most notable trait in human intelligence: the ability to create similarities and use things for various purposes, to develop common sense and have countless possible ways to face and resolve problems. Human intelligence allows us to understand a fable and its double meaning, going beyond a literal interpretation of the words. And even though we may not become Go masters, we can learn how to play it without needing millions of examples. Lee Sedol lost against AlphaGo, but what is striking is that he was even able to put up a fight, having played in his lifetime next to nothing, in terms of number of games, as compared to the computer. These examples demonstrate human superiority and bring into focus what machines have yet to achieve: the ability to have general intelligence, to abstract and transfer knowledge. A machine that can interpret images cannot recognize speech, while any two-year-old child can already do both. For an amateur chess player, it would be relatively easy to learn an alternative game where the rules are changed; for example, playing to capture the opponent's pawns. However, a computer's algorithms, despite being able to defeat the best chess player, needs countless new examples and training to play according to new rules. This is far from real-life scenarios in which we keep facing new situations, scenarios we have never experienced before and for which we have not been previously trained, but where we still know how to react based on common sense. AI advances in the last years have been truly astonishing, but human intelligence is still superior to that of a machine. We don't need to train with millions of examples for each task we solve, because we are the result of millions of years of evolution.

---

* SNARC network image courtesy of Gregory Loan.

* A thought cannot interact with matter unless we believe that, relying on exceptional qualities, we can manage to move objects at a distance or transmit thoughts like a Jedi Master. These paranormal skills, particularly telekinesis and telepathy, do not exist. The acclaimed magician James Randi dedicated himself to unmasking those who claimed to have these powers and offered one million dollars to whoever could demonstrate paranormal skills. Randi was not a multimillionaire, but he was sure that no one would ever be able to claim the prize. Among others, Randi publicly humiliated then famous Israeli "mentalist" Uri Geller, who claimed he could bend objects with his mind or read thoughts. Randi also showed how an evangelist who stated he could guess and cure the ailments afflicting his audience in fact was receiving radio signals from his wife, who was passing on information based on the prayers provided by audience members prior to the function. What was truly surprising was that these charlatans, regardless of how exposed they are, can continue to make a fortune through donations and book sales.

In Argentina, there have been endless swindlers who claimed to have psychic powers. In his crusade to expose these scams, TV presenter Raúl Portal showed how two Venezuelan "telepaths" shared information through a simple code hidden in their usual idle chatter. For example, to communicate to his accomplice what a person had drawn, one of them would quickly say, "*Concéntrate, ahora, rápido, ahora . . .*" (which literally means concentrate, now, quickly, now), and his blindfolded accomplice did nothing more than connect the first letters of each word to miraculously guess that the drawing was of a *cara*—a face.

* The problem for Tan wasn't his language comprehension but rather his articulation in a fluid discourse (both oral and written), despite not presenting difficulties in the movement of his tongue, lips, mouth, vocal cords, etc. In fact, he could whistle or hum a tune without any issues. Soon after, in 1879, the German neurologist Carl Wernicke would describe the opposite phenomenon: patients with lesions in the left temporal lobe who could talk (although only incoherent phrases) without understanding oral or written language. Language, then, is processed through two connected areas, what is known as *Wernicke's area*, in charge of language comprehension, and *Broca's area*, in charge of its execution. Lesions in Wernicke's and Broca's areas give rise to two different pathologies called *sensory aphasia* and *motor aphasia*. Patients with motor aphasia (Broca's aphasia) can barely pronounce isolated words, and those with sensory aphasia (Wernicke's aphasia) have no problem speaking, but what they say doesn't make sense. In some cases the lesion covers both areas, producing what is known as *mixed aphasia*.

* Together with James Watson and Maurice Wilkins, Francis Crick received the Nobel Prize in Physiology or Medicine in 1962 for discovering the double helix structure of DNA. During the last decades of his life, he devoted his time to studying the brain and the problem of consciousness, mainly with Christof Koch (who was my mentor at the California Institute of Technology).

CHAPTER 2

BLADE RUNNER

Can Androids Feel?

Cantela Insua

In *Blade Runner*, Ridley Scott portrays a Los Angeles of the future, with a labyrinth of big corporations housed in imposing buildings, in a perpetual night splashed with neon reflections and gigantic billboards. The action takes place in November 2019, a faraway future when the movie was released in 1982. However, Scott's vision does not differ much from the current panorama of some parts of Tokyo.

Based on the 1968 sci-fi classic written by Philip K. Dick, *Do Androids Dream of Electric Sheep?*, the plot is a futurist film noir in which Rick Deckard, an android hunter—or "Blade Runner"—pursues a group of androids (called *replicants*) who have entered Earth illegally. In the future imagined by Dick, replicants are used as a slave workforce for dangerous jobs in colonies on other planets, and are prohibited from coming to Earth under penalty of death, which in their case is referred to as *retirement*. These androids were genetically designed to be identical to humans, with one exception: they do not have true human feelings. However, over time, some began to develop their own emotional responses, becoming unpredictable—as a safeguard, a genetic mechanism was added to limit the replicants' lifespan to four years.

This fact launches the movie's plot. Cognizant of their relatively short existence, a group of replicants escapes to Earth to confront their creator and demand more years of life. But the underlying theme is much more interesting. The movie, like Dick's book, ponders the essence of what it means to be human and leads us to wonder what would truly differentiate a human from an android. Perhaps the most memorable part of the movie is when Roy Batty, leader of the rogue replicants, delivers his last words:

> **Batty:** I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I watched C-beams glitter in the dark near the Tannhäuser Gate. All those moments will be lost in time like tears in rain. Time to die.
>
> —*Blade Runner*

Batty's memories differentiate him from other replicants. These memories make him feel like a person despite not being human, justifying his quest to cling to a short life and, eventually, trying to extend it. After watching him die, Deckard reflects that Batty was looking for answers to questions we all ask ourselves: Where do we come from? Where are we going? How long do we have?* The separation between androids and humans is so vague in *Blade Runner* that Deckard ends up falling in love with an android, and movie fans still wonder if he wasn't a replicant himself.

The story's underlying discussion is closely related to a thought experiment that has led to intense debates in contemporary philosophy: the famous *philosophical zombie*. The p-zombie argument is this: Imagine that we could replicate a person down to the last detail, reproducing their brain, with all its neurons and connections among them. Would this replica be a sentient being, or would it be a zombie behaving like a human but unaware of its own existence?

We can think about this idea in the context of *Blade Runner*. Imagine if in the near future, humans discover unlimited amounts of uranium on a neighboring planet. The mining of such a reserve would mean the solution to all the energy problems on Earth. However, human miners would survive only a few years given that the levels of radiation on that planet are toxic.

The solution, then, would be to send replicants to do this job. Lacking consciousness, they are no more than machines, so we would have no problem sending them to a certain death. For this to work, we would need the replicants to be zombies; we would need them not to be conscious, because if they were, they would be sentient beings condemned to an unfair death. Moreover, the replicants—like in *Blade Runner*—would rebel against such a fate and fight to be treated as humans.

Is it then possible to replicate a person and succeed in having it behave like a human, but without being conscious? Some contemporary philosophers, such as Australian David Chalmers, argue that such zombies are initially possible because it is possible to conceive them; everything that is conceivable is possible; and, therefore, zombies are possible.[1] However, if we agree with Chalmers, we would have to ask ourselves what differentiates us from zombies or, in other words, what makes us conscious.

In the British TV series *Humans*, released in 2015, groups of androids assist humans in routine work, from manufacturing jobs to household chores like cleaning, ordering, or preparing meals. These androids are the perfect help because they lack consciousness and only need to recharge their batteries to continue working nonstop, without even asking themselves what they are doing and why. The action takes off when some of these androids become conscious and start to question their fates and the reason for their existence. The androids' "awakening" happens after they are fitted with a complex algorithm. But what would that algorithm look like? What would it have to do to make a zombie conscious?

If we accidentally hammer a finger, we scream in pain, but the pain is created by the activation of a specific group of neurons via the transmission of information through nerves called *C-nerve fibers*. A brief dialogue in the second part of *Terminator 2* illustrates this problem. Staring at the Terminator's bullet-ridden leather jacket in awe, young John Connor asks him, "Does it hurt when you get shot?" to which Terminator replies, "I sense injuries. The data could be called pain." In this context, we can wonder how the activation of specific neurons and nerves generates the sense of pain and, ultimately, why neurons activated when we hammer a finger make us feel pain and not pleasure, sadness, or tickles? In neuroscience, we know that when we see a red image, that image activates neurons in the visual cortex called *V4*.[2] Some of these neurons will respond to the color red, others to blue, and others to yellow, but this does not explain how our sense of red is generated—as in the experience we have when seeing a strawberry, a rose, or a bullfighter's cape. This subjective experience is what is known as *qualia*.

Reframing the question asked above, we scientists still discuss how certain physical and material processes—be it the impact of light waves on the retina or the activation of a group of neurons—lead to subjective experiences. This is the *explanatory gap*, or what Chalmers called the *hard problem*, the problem that he believes doesn't have, and will never have, a solution.

Chalmers's pessimism is not completely unfounded. After all, we are dealing with our own consciousness, and we lack an objective point of view. We can understand why an apple falls to the ground, why a dog salivates when hearing a bell that precedes food, but how do we understand what makes us understand these processes? How do we step out of the inherent subjectivity of our own thoughts to objectively evaluate what happens to us?

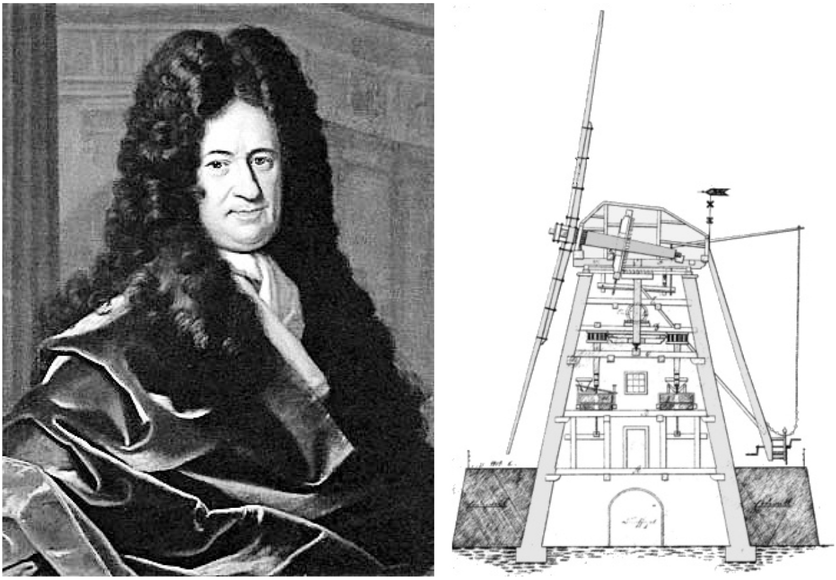
Perhaps we are not intelligent enough to understand ourselves. Although it is worth mentioning that the fact that we cannot understand something does not mean that the thing does not exist.

Gottfried Leibniz (1646–1716), the brilliant mathematician and philosopher who developed (independently from Sir Isaac Newton) the principles of differential calculus, was already asking himself these questions in the early eighteenth century:

> Moreover, we must confess that perception and what depends on it is inexplicable by mechanical reasons, in other words, through figures and movements. And if we imagine that there is a machine whose structure makes it think, feel, perceive, we could conceive it as enlarged, retaining the same proportions, in such a way that we could enter it as if it were a mill. Assuming this, when visiting it inside, we would only find pieces that push one another, and never anything with which to explain perception.
>
> —Gottfried Leibniz, *Monadology*

What Leibniz claimed about perception can also be said about consciousness. But contrary to Leibniz and Chalmers’s bleak outlook, many contemporary scientists believe it is possible to eventually understand consciousness, even though this may for many years continue to be the greatest challenge faced by science.

In fact, until a few decades ago we were asking ourselves what makes something alive. What makes a fly, an ombú, or an amoeba have life, but not a rock? In the early twentieth century, French philosopher Henri Bergson proposed the existence of an élan vital, a vital force or impulse of life, that is passed on from generation to generation in living beings. Today we know that such a mysterious substance does not exist, and that what is transmitted from generation to generation is deoxyribonucleic acid, or DNA. Similarly, we can ask ourselves if consciousness is something mysterious and inaccessible to human understanding, or if one day we will be able to explain it through scientific principles, just as was the case with the notion of élan vital.



Gottfried Leibniz and a windmill mechanism, which is used to defy the possibility of understanding how the brain works.

Let’s go back to *Blade Runner*. Imagine facing the group of replicants led by Roy Batty who demand to know why they aren’t treated like humans, why they’re denied access to Earth, and why they are exploited in unhealthy jobs. At first, we could argue that they aren’t treated like humans because they don’t have consciousness, because they don’t have minds. Yet by wanting to differentiate ourselves through the existence of our minds, we would be falling prey to Cartesian dualism—again, a trap that is hard to avoid because the dualist view of the body and mind is deeply rooted in our thoughts—and, as Daniel Dennett, perhaps the most brilliant philosopher of our time, says: Accepting Cartesian dualism is giving up.[3]

Most contemporary scientists endorse materialism—the idea that the mind is nothing more than the activity of neurons.* So, if we replicate one person’s neurons, we also replicate his mind, and the clone wouldn’t be a zombie roaming about but rather a conscious being indistinguishable from other people. But by embracing materialism, we face Chalmers’s “hard problem”: understanding how the activity of neurons can generate consciousness.

It should be noted that one of the most important obstacles in the study of consciousness is a strictly semantic problem. Since we don’t have a precise definition of what consciousness is, we tend to vaguely use the same word to describe different processes that are related but are not the same. In fact, the consciousness we have of our own existence, the one that shuts down when we sleep, is not the same as the consciousness we have while looking at a flower or thinking about Pythagoras’s theorem.

To avoid these problems, Francis Crick and Christof Koch proposed a radical paradigm shift in the early 1990s.[4] Until then, arguments about consciousness were the nearly exclusive domain of philosophers, and Crick and Koch’s proposal was to set aside discussions about definitions and classifications of consciousness (after several centuries, philosophers still haven't reached agreement), to also set aside universal solutions that explain everything, and instead foster the development of experimental paradigms that answer at least different aspects related to consciousness. In other words, instead of focusing on the hard problem, the idea is to experimentally explore simpler problems related to consciousness and, once some basic knowledge has been acquired, then go back to the big question. That means, instead of asking what consciousness is or how we can be conscious (something we will talk about later), we ask ourselves, for example: What happens in the brain when we recognize an image, and what is the difference when we don't recognize it?

One of the most used paradigms to experimentally address this last question is *binocular rivalry*. The idea is very simple. Our eyes are exposed to the same field of vision, although slightly displaced, given that they are separated by a few centimeters. So, 3-D movies don't do more than project, through special glasses, slightly displaced images to each eye, and that's why things seem to pop out of the screen. However, since our birth, both eyes have always seen the same images with a minimum shift.

What would then happen if, using 3-D glasses, instead of projecting the same displaced image to each eye, we project two completely different images? The answer is that instead of seeing a fusion of both images, we see one or the other, spontaneously alternating every few seconds.*[5] This principle is ideal to test consciousness. The key is that the stimulus hitting the eye does not change (what each eye sees remains the same), while the subjective feeling does, meaning the perception of seeing one image or the other alternates in time.

Now we can ask, what will the neurons do? Will they always respond the same way, since it is the same visual stimulus, or will they adjust their activity according to the person's perception? This is exactly the question that Nikos Logothetis—a brilliant scientist, director of the Max Planck Institute in Tübingen—and his team members asked themselves while recording the activity of neurons in the visual cortex of monkeys. Summarizing many years of work in one phrase, they found that most neurons in the primary visual cortex (the first area in the cerebral cortex that processes visual stimuli) did not change their activity, while most neurons in higher visual areas did, following the monkeys' subjective perception.[6]

Logothetis's work provides solid experimental evidence that goes straight to the root of the famous hard problem of consciousness. How can a physical stimulus give way to subjective feelings? Easy: neurons in the primary visual areas represent the basic characteristics of the stimulus, and after subsequent processing stages, a representation of the subjective feeling is generated in higher areas. The different levels of processing form an increasingly abstract representation to extract a meaning from the stimulus.* It is no coincidence that in the previous chapter we observed that the subsequent layers of Hinton's deep networks generated an analogous representation of the images analyzed. The design of such networks is precisely based on the principles of visual processing used by the brain.

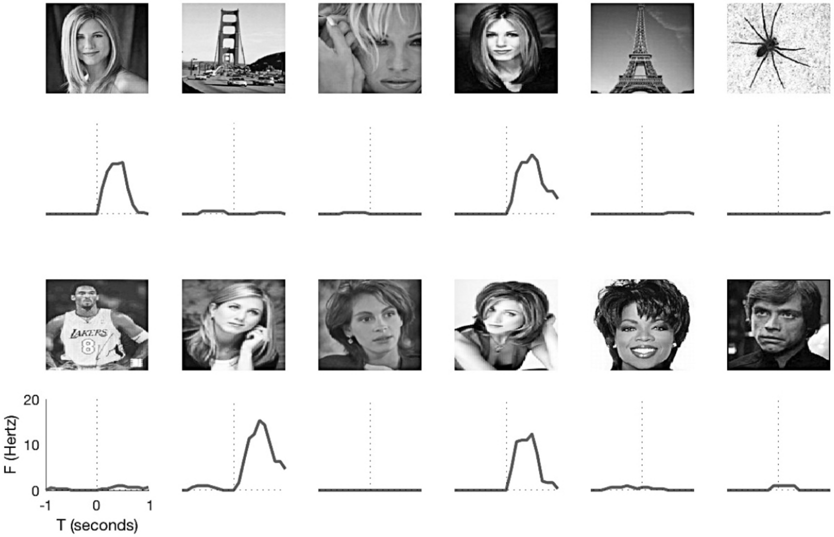
The higher visual areas project to the hippocampus, the seahorse-shaped structure that was extracted from patient H.M. to try to cure his epilepsy. We saw that after the surgery, H.M. lost the ability to make new memories. Based mainly on the information provided by H.M. and by subsequent lesion studies in animals,[7] today no one questions the key role that the hippocampus has in memory. Yet, surprisingly, the surgery that had disastrous results for H.M. nowadays is frequently performed and with pretty high chances of curing the patient.**

Most patients with epilepsy can be treated with medication, but about 20 percent of the time medication doesn't inhibit the seizures, and these patients are candidates for a surgical resection of the *epileptic focus*. Often, these resections include the hippocampus because it tends to be involved in the origin of epileptic seizures. However, would we be condemning the patients to end up amnesic like H.M. by doing this? Of course not. The big difference is that only one of the hippocampi is extracted, the left or the right (and some neighboring areas), according to the case, but never both. That said, if surgically removing an area of a patient's brain is being considered, it is vital to ascertain that this is the area that triggers the epileptic seizures. For this reason, in many cases surgeons implant electrodes in the brain to record activity during epileptic seizures and accurately determine the location of the focus.

I reiterate that the reason these electrodes are implanted—as well as the decision of how many electrodes are placed in each patient and in what areas—is exclusively clinical; the objective is to cure the patient. Although, on the other hand, such a procedure allows us to study the behavior of neurons in the human brain.[8] The unique opportunity to perform these recordings is what convinced me to pursue a postdoctoral degree at the California Institute of Technology (Caltech) and what led me to the discovery that radically changed my scientific career.

Implantation of intracranial electrodes in a patient with epilepsy.

In the first experiments, we showed the patients images of actors, politicians, athletes, famous places, and so on, while we recorded the activity of dozens of neurons.[9] The idea was to see if any of these neurons responded to any of the images. Setting aside countless technical details, the results surpassed by far the most optimistic expectations I had when I first moved to Los Angeles to work on this project. We found neurons in the human hippocampus that responded to concepts![10] The most famous of these neurons was one that responded to different pictures of Jennifer Aniston and not to photos of many other people, including other actresses such as Julia Roberts, Pamela Anderson, and so on.



Responses of one neuron in the hippocampus that only reacted to different pictures of Jennifer Aniston. The thick lines indicate the average response to six presentations of each picture, shown for one second, starting at time zero. (Due to space considerations, only four of the seven pictures of Aniston and eight of the other eighty pictures used are shown here.)

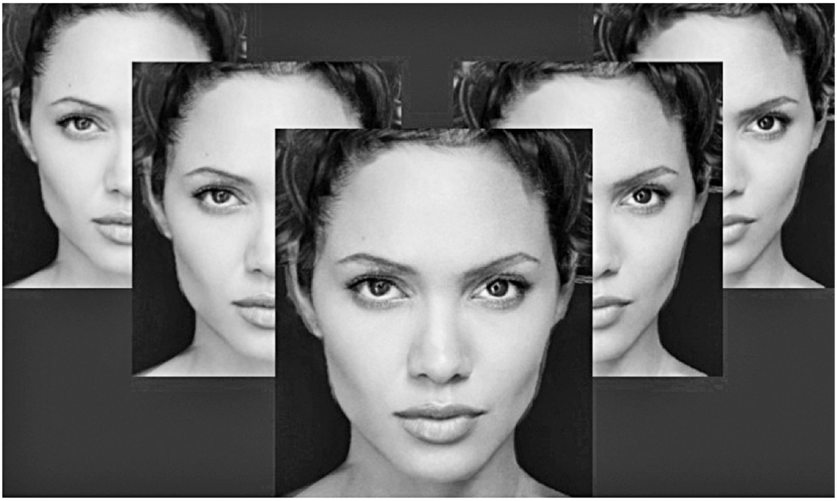
Likewise, we found another neuron that only responded to different photos of Halle Berry, another that did so with Oprah Winfrey, another that responded to the character Luke Skywalker, and dozens of other examples.[11] The fact that these neurons became active with different pictures of one person but not others implies that they respond to the person, the concept, rather than the particular details of one or another picture. This was the most surprising finding for my colleagues and, of course, for me too: the incredible degree of abstraction, of representation of concepts, by these neurons. Furthermore, the neurons also became active when we showed the name of the person that they responded to—and not when we showed other names—and even when the name was pronounced by a computer-synthesized voice.[12]

I have taken another big detour to finally be able to explain a couple of experiments that are closely linked to our discussion about consciousness. Suppose we show Jennifer Aniston's picture to a subject but instead of leaving it on the screen for a second, such as in the first experiment, we flash it very briefly (for less than one-tenth of a second, to be more precise). The idea is that, by showing the photo so quickly, the subject sometimes recognizes the person and sometimes does not. Let's also suppose that we do this while recording the activity of the neuron that responds to Jennifer Aniston. What will this neuron do? Will it respond each time we show her picture—after all, the visual stimulus is exactly the same—or only when the subject recognizes Aniston? This is exactly the idea behind the experiment we did a few years back, where we proved that if the subject recognized the picture, even if it was shown for a small fraction of a second, the neuron more or less responded in the same way as when the picture remained on the screen for a much longer time. On the contrary, if the subject alleged having not seen the picture, the neuron remained silent, as if we hadn't shown anything. In other words, these neurons reflected the subject's conscious perception and not the visual stimulus, which was always the same.13

In a later experiment, we continued exploring the same idea but used another task. Instead of quickly showing a picture for the subject to sometimes see and sometimes not, we morphed two different faces. For example, starting with a picture of Angelina Jolie and another of Halle Berry, we created varying degrees of morphing between the faces and evaluated which of those was the most ambiguous, that is, which one the subject sometimes saw as one person (Jolie) and sometimes as the other (Berry). Then, if we determined that a neuron responded to Halle Berry, we examined what this neuron would do when we showed the ambiguous morphed image of Berry and Jolie. Would it respond in the same way, since the photo was the same, or would it only respond when the subject said it was Halle Berry?

The latter was the case: neurons fired only when subjects said the morphed image was one person—in fact, the neurons responded just like when we showed the original picture—and did not fire if, when seeing the same image, the subjects said it was the other person.14

Summarizing the previous results, in the primary visual cortex there is a representation of the physical stimulus, while in higher visual areas there is a representation of the person's subjective perception. The higher visual areas, in turn, send information to the hippocampus, where the representation is much more abstract and tied to the subject's consciousness. It doesn't matter if what we show is the person's picture or their written or spoken name. It doesn't matter if we barely show a glimpse of a picture or if we use a morphed image. When the subject recognizes the person, regardless of the stimuli used, the corresponding hippocampal neurons will fire.



Morphed images used to test the responses of hippocampal neurons. The left picture is of Angelina Jolie, the right one is Halle Berry, and the middle one is an ambiguous morphing between both of them, which subjects sometimes identified as Jolie and other times as Berry.

Are hippocampal neurons involved in consciousness? This is a tricky question because it is very vague. Again, it depends on what we mean by consciousness. If by consciousness we refer to the recognition of something, that is, of being conscious that we are seeing Halle Berry and not Angelina Jolie, then the answer is a categorical no. Why? Because patient H.M. had both hippocampi removed but could still recognize his mother, the surgeon who operated on him, the house where he lived, and other people and things that he knew prior to the surgery. H.M. couldn't create new memories, but his perception was normal.[15] The process of recognizing Halle Berry's picture is not the same as the one that brings to consciousness an avalanche of memories related to her: "I remember seeing her in a James Bond movie," "I remember finding a neuron that responded to her," "I remember using images in which I morphed her with Angelina Jolie," and so on. And the memories we bring to our consciousness can be diverse: When I see a photo of jasmine, I remember its fragrance, the texture of its petals, the jasmine plant at my grandmother's house in Chacabuco. All these memories are part of my subjective experience upon seeing jasmine. The sum of all of these feelings and memories is, as described above, what we call qualia.

Consciousness, then, involves at least two different processes (and probably many more): on one hand, the process we use to recognize a person, and, on the other, the process we use to bring to consciousness the information related to that person.[16] What would happen if we didn't have hippocampi, if we didn't have the neurons I've just described? Eventually we could recognize objects, animals, and people, but we probably wouldn't remember the experiences related to them. Such was H.M.'s postsurgical state of permanent confusion, of living in an eternal present with no access to the past. In fact, H.M. used to say that "each day is itself unique, whatever the joys or sorrows I may have had."

Leonard Shelby, the main character of the neo-noir thriller *Memento* (released in 2000), provides a spectacular portrayal of a case similar to that of H.M., and what it means to live without the ability to form new memories. To remember something related to a person, Leonard wrote notes on a Polaroid photo of that person. The movie revolves around the dramatic decisions he makes based on these quickly sketched notes, as if they were his only source of long-lasting memory. Neuropsychologist Brenda Milner, who studied H.M. for thirty years,* described his state as that of someone who is constantly beginning to understand his circumstances but without fully understanding what is happening because he can't put his present situation in context with his immediate past. H.M. said, in that regard, that his life was like waking from a dream.

We usually wake up in our beds and everything is familiar, but if we spend the night in a hotel, when we wake up, we tend to have that brief moment of confusion when we don't understand where we are or what we are doing there. However, our memory quickly saves us by reminding us of the circumstances that led to us being in that place, and everything makes sense again. In H.M.'s case, that moment of confusion was endlessly extended because he couldn't remember where he was or what he was doing wherever he was. The first scene in *Memento* brilliantly illustrates this idea. It begins with Leonard's voice-over while waking in an unknown place:

> **Leonard:** So, where are you? You're in some motel room. You just wake up and you're in . . . in a motel room. There's the key. It feels like maybe it's just the first time you've been here but . . . perhaps you've been there for a week . . . three months, it's . . . it's kind of hard to say. I . . . I don't know. It's just an anonymous room.
>
> —*Memento*

H.M.'s case may perhaps be the closest one to the philosophers' zombie. We have already established that such zombies couldn't actually exist; if we replicated the human brain, we would also replicate its consciousness. But if in the cloning process there was a glitch when reproducing the hippocampus, the being we would create would at first behave like other humans (despite his state of permanent confusion, H.M. could recognize things, hold conversations, and even had a better than average IQ), but the clone would have memory issues and a very limited consciousness of itself and its circumstances (and note that I do not regard consciousness as something that either exists or doesn't, but rather as something that can be present at different levels). Oddly enough, when H.M. was fifty-five years old, he didn't recognize himself in a picture taken fifteen years earlier.* He also wasn't aware of the passage of time and events. For example, in April 1955, a year and a half after his surgery, he was barely aware of having had surgery, and he calculated that the date was March 1953 and that he was twenty-seven years old, when he had already turned twenty-nine.

We have seen how the brain generates subjective feelings from a representation of physical stimuli. We have also seen that we already understand, at least partially, how what we call consciousness is generated. It might seem that we have answered the big question, that we have solved the hard problem. Yet, despite having made great advances in recent years, the problem apparently still stands. After all, we haven't explained how neurons in the hippocampus, or in higher visual areas, generate our subjective feelings. We would have to explain, for example, why we are conscious of the activity of these neurons and not of the activity of many others that regulate the secretion of hormones or the diameter of our pupils. What makes us conscious of the activity of some neurons and not of others? In the end, what leads us to feel something? These questions continue to spawn big debates, and we will revisit them later in this book.

In chapter one, we discussed the astonishing advances of artificial intelligence in recent years, especially since the use of deep neural networks. We argued that the capability of consciousness is not limited to organic beings, and it could, in principle, also exist in sufficiently sophisticated circuits and algorithms. But how could we know this? How could we test if a machine is conscious, if it is effectively able to think by itself, or if it is limited to giving sophisticated answers based on complex algorithms?

In *Blade Runner*, Rick Deckard examines potential androids using a Voight-Kampff machine, which measures unconscious physiological responses—like the dilation of pupils—while asking emotionally loaded questions. The idea is to see if the person being interrogated can maintain a coherent conversation about their past and show physiological responses that, lacking feelings and emotions, replicants do not have. In the world of the story, new generations of replicants could, however, incorporate the ability to reproduce such responses. In fact, current real-world androids, like the Geminoid or Actroid models, are close to reproducing human gestures. However, the most difficult task for these androids would be to be able to maintain a coherent conversation and interaction, knowing when and how to react. This is exactly the idea of the Turing test, proposed by English mathematician Alan Turing in 1950.

Turing argued that asking if a machine is able to think is like asking if a machine can replicate human behavior. He proposed what he called the *imitation game*, where an evaluator interacts simultaneously with a real person and with a computer in another room, and tries to determine which is the machine and which is the human.[17] The interaction takes place through a monitor and a keyboard, to set aside technical cues such as the tone of voice, the robot's appearance, and so on. If, based on this online chat, the evaluator cannot distinguish between the two, the machine has passed the test.

The Turing test has been used since the 1990s, in the (discredited) Loebner Prize, an annual competition that offers a couple thousand dollars as an award to the algorithm that best replicates a human conversation. This competition has been so discredited—after too many crude publicity stunts without any scientific backing—that Marvin Minsky, the father of AI introduced in chapter one, once offered a monetary compensation to whoever managed to abolish this prize.

Aside from this criticism, the Loebner Prize has been a colorful forum presenting different algorithms—called *chatbots* (an abbreviation of *chatting robots*)—that interact with users. For example, Mitsuku,[18] one of the chatbots that won this competition, provides a pleasant and reasonably coherent conversation, so long as the questions are relatively simple (and it conveniently tends to change the subject if it doesn't understand the question).* Principles like those implemented in chatbots are increasingly used every day in applications like Siri, which allows verbal interactions with Apple iPhones and iPads, offering the services of a (for now) very limited, yet efficient, assistant that manages emails, calendars, internet searches, and other basic tasks.

Getting back to the Turing test, the idea seems to make sense since we can imagine endless questions with which we, like Rick Deckard, could detect whether we are interacting with a machine. But, in practice, the results are controversial because they are deeply influenced by the evaluator's efficiency in formulating adequate questions to detect the chatbot. It would be unfair to blame Turing for this. After all, he was trying to propose a practical way of determining if a machine can think, avoiding endless discussions about what is understood by *thought*. It is very likely that Turing himself would accept that nowadays there are other, more efficient ways to approach this question. In chapter one, we saw that algorithms based on deep networks are capable of recognizing and even interpreting the content of different photos. This achievement is much more relevant than if a chatbot can confuse an evaluator with whom it interacts.

Generally, chatbots can't answer complex questions or those with double meanings. However, in February 2011, an IBM supercomputer called Watson crushed two historic *Jeopardy* champions. What's surprising is that in this game, questions aren't explicitly formulated—such as, "Who is . . .?" or "What is . . .?" and so on—but are instead implicit in relatively intricate texts that use double entendre and information in different contexts. Just understanding the meaning of the actual questions is by itself a great challenge for any computer, but Watson not only understood the questions but also gave the correct answers (using its memory database without an internet connection) quicker than the other two, until then, unbeatable contestants.

Another known demonstration of AI's achievements happened in 2012. Rick Rashid, director of Microsoft Research, spoke about automatic speech recognition at a conference in China. While he addressed the audience, a recognition algorithm based on deep networks wrote everything he was saying on a gigantic screen. This was only the beginning. After explaining technical issues involved in speech recognition and the latest advances in this field, Rashid went on to talk about Microsoft's progress in the automatic translation of texts. He explained how it was possible first, for an algorithm to recognize what he was saying (as shown on the screen); second, for another algorithm to translate it into Chinese (as shown on another screen); and finally, how it was also possible for a third algorithm to pronounce those words with Rashid's own voice. To the audience's amazement, Rashid finished his conference speaking in Chinese via a synthesized voice that translated his words as soon as he pronounced them.[19] I specifically mention this demonstration because it is the perfect illustration of a discussion that has been going on for years in philosophy: the famous *Chinese room argument*.

In addition to the aforementioned limitations of the Turing test, a much more substantial criticism of AI was made by philosopher John Searle, who maintained that the Turing test in fact could never establish if a machine is capable of thinking. To illustrate this point, Searle proposed a thought experiment: Imagine a person in a room with a large instruction manual on how to operate Chinese symbols. The person receives, from the outside, cards with questions in Chinese that he does not understand but is capable of answering by following the instruction manual.

Searle's conclusion is that this person would appear to understand Chinese despite not speaking a word and would pass the Turing test. Searle uses this argument to deny the possibility that a machine is capable of thinking, since, according to him, it would be like the case of the person in the Chinese room, who can only follow rules without understanding their content. When hearing about Watson's achievement on *Jeopardy*, Searle congratulated the IBM scientists but concluded that Watson didn't even realize it had won.

By the way, not everyone shares Searle's pessimistic view on AI.[20] After all, what makes us, and not a machine, able to understand? What's the difference, and why can't this difference be implemented in a machine?

The vast majority of the scientific community agrees that the question of how to determine if a machine is capable of thinking and being conscious still has no answer. We have described the Turing test, but also the criticism of both the underlying idea (the discussion about the Chinese room) and its practical implementation (the Loebner Prize). In my opinion, the person in the Chinese room could demonstrate that they understand Chinese if they could show the ability to generalize, that is, to answer at least some questions without following the rules in the instruction manual.

Of course, Searle would argue that the manual covers every possible question—and that we don't have access to see if the person consults the manual or not—but this is different from the posit we made about AI. A deep network developed to recognize images isn't programmed, or rather, trained, to play chess. So, we could argue that a machine is capable of thinking if it can solve problems for which it wasn't explicitly programmed (or trained with endless examples), based on generalizing and transferring knowledge.

We ended the last chapter with Marvin Minsky's ideas about the ability, up to now exclusively human, to generalize and use common sense. Minsky highlighted the importance of using analogies in human thought, contrary to the functioning of machines, which is based on the implementation of logical rules—if *A* happens, I do *B*. Making analogies is precisely the basis for transferring knowledge, to apply what has been learned in one task to another, completely different task. Remember how easy it is for an amateur chess player to adapt to changes in the rules, such as playing to capture the opponent's pawns. The player understands the instruction and transfers knowledge without needing much practice, while a computer needs to be retrained from scratch.

The human ability to make analogies and generalize is based on extracting the meaning of things. The extraction of meaning depends, in turn, on being able to make abstractions, keeping the essential and discarding the rest. When riding a camel, I transfer knowledge I acquired riding a horse, and I adapt to the peculiarities of the camel. But it doesn't matter if I learned with a white, brown, or jet-black horse. In this context, I set aside all of those details and extract the information on how I balanced myself on the horse to apply that knowledge to riding the camel. This is also the foundation of our creativity, of making analogies and associations between facts that apparently have no connection. Newton's brilliance was not the discovery of the exact formula of the law of gravitation, but in the fact that he realized that disparate things, like an apple falling from a tree and the moon remaining suspended in the sky, respond to the same phenomenon. It doesn't matter if the apple is big or small, red or green, just like it doesn't matter if the moon is at the zenith, or if it is waning or waxing. All that matters is that the apple is a static object and that's why it falls, while the moon is an object in orbit and, therefore, it continues on its course.

Compared to the storage capacity of a computer, humans remember very little. More than keeping a reliable record of everything, our brain seeks understanding. The processes of abstraction and attribution of meaning are not easy to implement in a computer—since deciding what information to use and what to set aside is not obvious—and these processes lead to a construction of analogies, to the transfer of knowledge, and to the development of general intelligence. Moreover, the process of abstraction generates internal and subjective representations. This is precisely what is coded by our concept cells: abstractions—concepts that are closely tied to our consciousness, since these neurons fire only when we are aware of seeing something. In other words, the absence of these processes in a computer is what makes it lack general intelligence and internal representations that, I boldly suggest, perhaps are the cornerstone of our thoughts and our consciousness.

---

* These thoughts, captured in a voice-over in the movie's original version, were eliminated in subsequent releases. While voice-overs are often characteristic of film noir, in this case the voice-over was added at the insistence of producers to make the plot easier to understand, something director Ridley Scott believed was unnecessary.

* To avoid overcomplicating things, I'm substantially simplifying this idea and do not delve into more refined philosophical distinctions. Basically, what I want to say is that almost no one in today's scientific world supports Cartesian dualism and the idea that the mind is something that goes beyond the activity of neurons.

* With Mariano Molina, a great painter and friend who spent a year in my lab, we used this principle to create paintings (obviously made by Mariano) that, when seen with 3-D glasses, change every few seconds. What's surprising is that the paintings were nothing more than fabrics on the wall, static objects that acquire movement when seen with 3-D glasses. These paintings, together with many others, were exhibited in England. For more information, visit doi.org/10.1016/S1474-4422(12)70055-8.

* Of course, there's a lot more to explain about how this process of abstraction and extraction of a meaning works, but I won't get into the details here because this is a recurring theme I developed in a previous book, *The Forgetting Machine* (BenBella Books, 2017).

** Although, unfortunately, not all cases can be surgically treated, since this depends on the location of the epileptic focus, the type of epilepsy, and many other factors. Since I am not a physician, it would be unethical for me to delve into more detailed clinical aspects related to epileptic surgery. My objective here is just to explain the reason why electrodes are implanted in the brain of patients with epilepsy.

* After H.M.'s death in 2008, Brenda Milner said she had a strange feeling, as if she'd lost a great friend, even though, after decades of constant interaction, that friend never managed to recognize her.

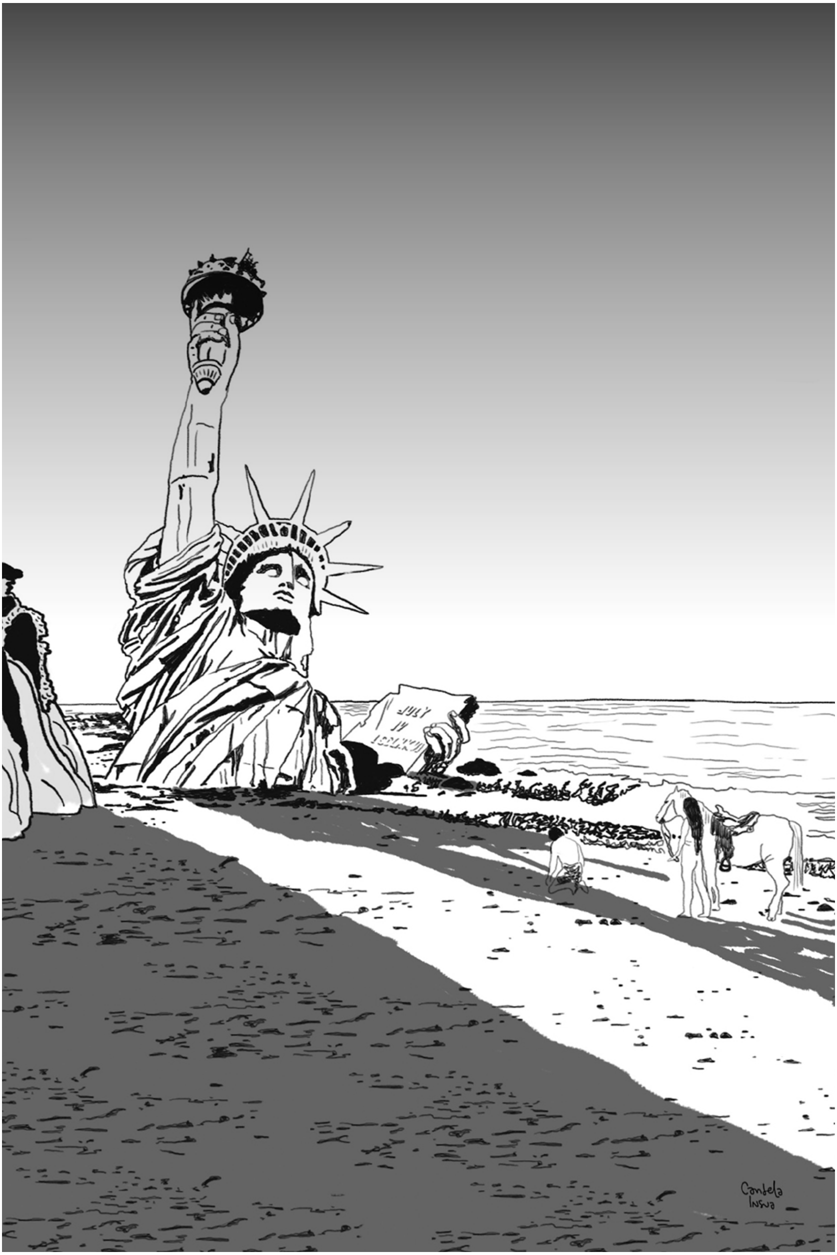
* In "The Lost Mariner" (a story featured in the book *The Man Who Mistook His Wife for a Hat*), Oliver Sacks writes a dramatic description of the confusion and panic that a patient with amnesia felt when he saw his face in the mirror, believing he was nineteen years old when he was close to turning fifty.

* Setting aside the differences, the interaction with these chatbots brings to mind Spike Jonze's 2013 movie *Her*, where the main character falls in love with a computer application with which he interacts.

CHAPTER 3

PLANET OF THE APES

Animal Consciousness


JULY
IV
Candela Insua

A few months before Kubrick blew everyone away with *2001: A Space Odyssey*, the movie *Planet of the Apes* was released, based on the book *La planète des singes* by French novelist Pierre Boulle. It is noteworthy that both movies premiered in 1968, highlighting the magnitude of *2001* compared with one of the biggest sci-fi blockbusters of its time. All you have to do is contrast the routine scenes of Dave Bowman in the *2001* spaceship with those on the spacecraft in *Planet of the Apes* to notice the glaring difference in production quality. This comparison may be unfair because *2001* could premiere tomorrow, more than a half-century later, and still be considered cutting-edge, with a cinematographic aesthetic and depth of content that never cease to amaze. Yet this assessment has nothing to do with how both movies were received in their time.

The reviews of *2001* were mixed, given its extremely complex story line, endless scenes in space, and almost total absence of dialogue. *Planet of the Apes*, on the other hand, was an immediate hit, based on a linear plot touching on highly relevant themes at the time: the Cold War boom and widespread fear that the human race would be annihilated (the movie launched only a few years after the Cuban Missile Crisis, which threatened to unleash a nuclear conflict between the United States and the Soviet Union); issues surrounding racial assimilation (civil rights activist Martin Luther King, Jr., was assassinated that same year); and heated controversies triggered by animal research.

In *Planet of the Apes*, a team of astronauts in suspended animation travels to an unknown planet. Given that the craft is moving at close to light speed, time goes by much slower, and while for the crew the trip lasts a year and a half, more than two thousand years have gone by on Earth. A malfunction leads them to crash-land on a seemingly deserted planet, where after a lengthy trek they encounter a tribe of primitive humans who don't have the ability to speak and are dominated by a much more advanced race of apes. Taylor, the astronauts' leader, is captured and studied by two ape scientists, who are astonished by his attempts to communicate (he is temporarily unable to speak due to an injury to his throat sustained during his capture). Perhaps the most notable scene is at the end of the movie, when Taylor, wandering aimlessly after escaping his captors, discovers the ruins of the Statue of Liberty and realizes he has been on Earth the whole time.

In that faraway future, the human race has been brutally overpowered after, as Taylor learns, having been nearly exterminated by some past war. The exercise proposed by the movie of inverting the evolutionary scale leads us to put ourselves in the "shoes" of an oppressed species and reflect on our interaction with animals, since they are very likely thinking and conscious beings like us, although less developed.

We previously explored the advances and limitations of AI and the possibility of thought and consciousness in machines and androids. *Planet of the Apes* triggers a much more concrete discussion. We are no longer talking about sophisticated algorithms or circuits we might eventually develop, but about beings that share the Earth with us. Can they think? Are they conscious of their own existence? Do they fear their death? In other words, do animals reflect on the things that happen to them, or do they just act mechanically, responding instinctively to different stimuli?

The Greek philosopher Aristotle (384–322 BC) differentiated three levels of intellect: The *vegetative soul*, possessed by plants and other living beings, is in charge of procreation and food. The *sensitive soul* covers the senses and is characteristic of animals because any animal has at the very least the sense of touch, while the *rational soul* allows for reasoning and is exclusive to humans. According to this view—discussed in his famous monograph *De Anima*, or *On the Soul*—the ability to reason is what distinguishes us from other animals, and from there stems the famous characterization of man as a "rational animal."

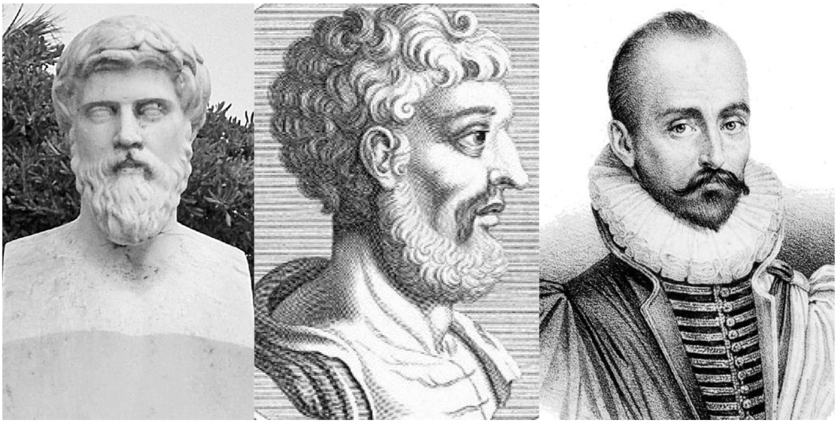
The Stoic philosophers of ancient Rome agreed that animals weren't gifted with *logos* (understanding) and behaved by following a natural survival instinct. They were seen as beings created for the benefit of man, who therefore had no need to feel moral obligation toward them. This characterization of animals was crucial for the Stoics. Since their philosophy was based on the development of virtue, if they started to doubt their view of animals as lacking rationality, they would have to reconsider the way they treated them as well as the very foundation of their doctrine.

A visceral criticism of this position was offered by Plutarch (46–120 AD) in *Moralia* (Matters Relating to Customs and Mores), a collection of seventy-eight essays where he covers a wide variety of subjects, from the inconvenience of being in debt, to the visible faces of the moon, to the existence of the soul according to Plato. In several of these essays, Plutarch aims to demonstrate the ability of animals to reason, arguing that they have intelligence, although not as developed as humans. Consider this excerpt from "On the Intelligence of Animals":

> Many animals, my friend, surpass all of humanity in size and speed, with their penetrating sight or fine ear, but that does not mean that man is blind or invalid or lacks ears; we also run, even if it is slower than deer, and we see, although worse than falcons; nature hasn't deprived us of strength or volume, even though in these aspects we are nothing when compared to an elephant or camel. So, if animals have clumsier mental activity or worse reasoning, neither should we say that they don't reason or have mental activity at all, nor that they lack reasoning, but rather that they possess a weak and murky logic, like an eye with limited and disturbed sight.
>
> —Plutarch, "De sollertia animalium" (*Moralia*)

Plutarch illustrates this idea by describing the behavior of several animals, including behavior observed in foxes that were used by the inhabitants of ancient Thrace to confirm the solidity of the ice before crossing a frozen river. Foxes move slowly and place their ear to the ground to estimate the thickness of a layer of ice by listening to the current flowing below. Likewise, he claims that a dog that must choose among three paths to follow another animal, when not finding traces in the first two paths, infers that its quarry has taken the third path. With these examples, Plutarch aims to demonstrate that animals not only use their senses but also their logic to make decisions.

Plutarch's ideas were picked up by the physician and philosopher Sextus Empiricus (160–210 AD) in *Outlines of Pyrrhonism*, a treatise that follows the Greek thinker Pyrrho's tradition of skepticism, questioning objective truths, and later by Michel de Montaigne (1533–1592) in *An Apology for Raymond Sebond*, where the French philosopher reconsiders, also with a big dose of skepticism, the dogmas of the Catholic Church. In this widely regarded essay, Montaigne develops an unwavering defense of animal reasoning, giving endless examples of their intelligence and famously asking himself: "When I play with my cat, how do I know that she is not playing with me?”



Plutarch *(left)*, Sextus Empiricus *(center)*, and Michel de Montaigne *(right)*, three famous defenders of animal intelligence.

Montaigne despised the arrogance and, specifically, the vanity with which humankind sees itself compared to other animals. He was very critical of the anthropocentric view of the universe proclaimed by Christianity, where humans are considered the chosen species with the unique ability of intelligence and thought. His words leave no doubt:

> [O]n what foundation has [man] built the staggering superiority that he believes to enjoy over other creatures. Who has taught him that the admirable motion of the celestial vault, the eternal light of those torches proudly held above his head, the tremendous tremors of that infinite sea, have been established and continue for centuries upon centuries for his convenience and service? Is it even possible to imagine anything so ridiculous as this miserable and emaciated creature who isn’t even master of himself, who is exposed to being harmed from all ends and that, nevertheless, believes himself to be the emperor and ruler of the universe, when he doesn’t even know the tiniest of its parts, let alone being able to govern it? . . .
>
> Why, then, do we consider them [animals] to be deprived of souls, life, and reason? Have we been able to recognize in them immobility and insensitivity, having no other relationship with them than one of submission and obedience? Do we dare say that we haven’t seen in any creature other than man the use of a rational soul?
>
> —Michel de Montaigne, *An Apology for Raymond Sebond*[1]

A few years after Montaigne’s death came the birth of the most brilliant French philosopher, René Descartes, who would provide a radically different view on the potential intelligence of animals. As we have discussed, Descartes postulated the division between body and mind: The body is in the physical domain, while the mind is exclusively human and it is the pillar of our thoughts. For this thinker, animals are like robots that seem to act as if they were conscious, but in reality they are not, since they only react mechanically based on reflexes, without the ability to reflect on their acts.

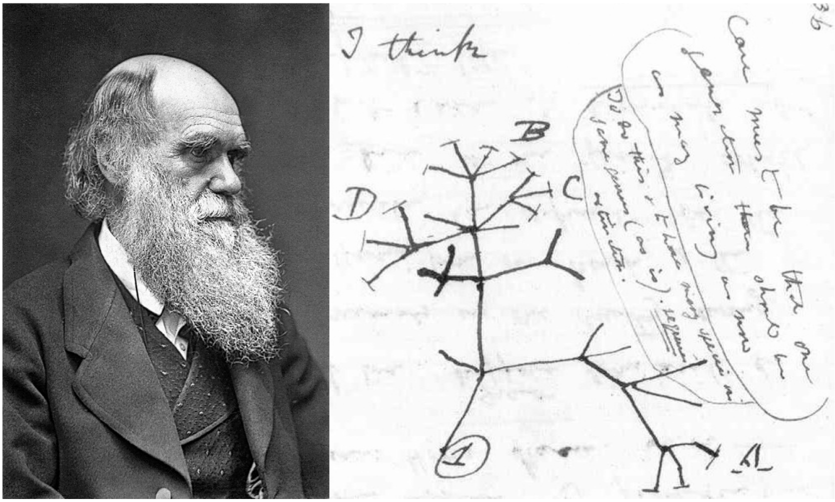
Like Aristotle, Descartes didn't deny that animals could feel, but he dismissed their capacity to reason. In his view, animal behavior and its perceptual ability can be explained through mechanical principles that do not require comprehension. In the fifth part of his monumental *Discourse on the Method*, he bases this position on two facts: First, animals lack an articulate language, which is the means to express thought, so he concludes that they can't show their thoughts simply because they do not have them. Second, animal behavior is limited to specific tasks, while human behavior is based on reason, a universal tool that can be used in different circumstances. (Note the relation to Minsky's arguments and our foray into general intelligence at the end of the first chapter.)

In the quote above, Plutarch argues that animals can even surpass humans in diverse tasks. Descartes rejects this argument, alleging that this does not prove they have wit since it is like saying that a clock, made of wheels and springs, can count the hours better than a person.[2] Montaigne, conversely, says that animals also communicate among themselves, even if we don't understand them, "just like we don't understand foreigners who come from other lands and we consider them to be ignorant because they don't know our customs and remain silent because they can't speak our language."[3] However, Descartes claims that animals have organs that are similar to ours, with which they could make themselves understood.[*] Notably, both Plutarch and Montaigne underscore the ability of magpies and parrots to speak, but Descartes notes that these animals can only repeat words without "speaking," meaning without being able to express thoughts.[**] Furthermore, Descartes adds that humans who are born deaf or mute, lacking the organs that allow them to speak, can still communicate using signs.

None of these arguments by Plutarch, Montaigne, and Descartes offers overwhelming proof, be it in favor or against the intelligence of animals. However, in the mid-nineteenth century, English naturalist Charles Darwin came up with the theory that would revolutionize our view of animals and the very roots of our origin. In his explosive tome *On the Origin of Species*, published in 1859, Darwin proposes that with the passage of time, from generation to generation, species go through changes (which we now attribute to DNA variations). Such changes aren't necessarily beneficial, but, given a process of natural selection, organisms will compete for the same resources—water, food, space—and the fittest will survive, while those with variations that are less adapted to the environment will become extinct. Of course, this process takes millions of years, but the final result is that, given the random mutations and the process of natural selection, species evolve with time.

From this theory comes the tenet that all species have common ancestors: simpler life forms that evolve, like the branches of a tree that come from the same trunk. With my closest relatives (my siblings), I have immediate common ancestors (my parents); with my cousins, the ancestors are further removed (my grandparents); and the less related my family members, the more distant the common ancestors. Similarly, the closest species, like man and higher primates, have relatively recent common ancestors on the evolutionary scale, while in order to find common ancestors with those species with which we have greater differences—for example, birds and reptiles—we must go further back in time.

Darwin's theory of evolution doesn't distinguish between humans and other animals. We all come from the evolution of different branches of the tree of life. The notion of evolution through natural selection, one of the simplest and yet most astounding ideas in the history of humanity, was an incredible blow to the very basis of Descartes's anthropocentric conception. If we all descend from the same ancestors, why are we the only ones capable of thinking and being conscious of our existence? Or, as Montaigne would have said, “And that privilege that man has attributed to himself in this proud building where he pretends to be unique regarding the ability to recognize beauty . . . who has given him such a privilege?”[4]

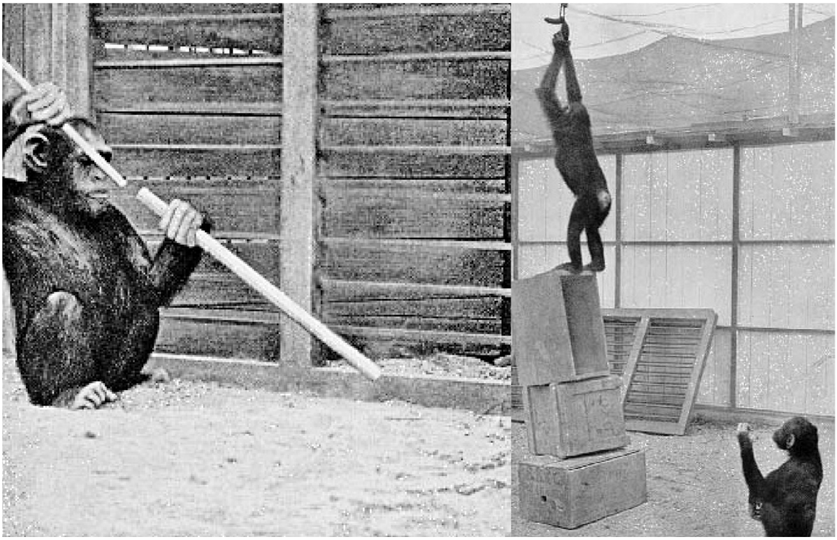


Charles Darwin and his illustration of the species evolution tree in a notebook in 1837. In the text on the right, Darwin says that “in one given generation there should be as many species as there are now and that this, in turn, requires extinction.”

DNA is the genetic code transmitted from generation to generation in living beings, a type of instruction manual that determines the development of cells in an organism. My siblings’ DNA is very similar to mine because we descend from the same parents and have common features we inherited from them. But no matter how similar two people’s DNA may be, it is always different, and these differences determine the specific characteristics that makes us individuals. On average, the DNA variation between any two people is around 0.1 percent. This difference makes some of us taller than others; gives us dark eyes; sometimes light, curly hair; a birthmark; or certain innate behaviors.

To give you an idea of the differences we have with other species, based on studies of prehistoric remains, we now know that the difference between our DNA and that of a Neanderthal—whom we were separated from half a million years ago on the biological tree—is around 0.5 percent. The separation from chimpanzees, bonobos, and gorillas, our closest relatives, happened twelve million years ago, and we share almost 99 percent of our DNA with them. With orangutans, the difference is a little bigger (around 3 percent); with the rest of the monkeys it is even bigger; and so on. Given the minimal difference between our genetic code and that of higher primates—chimpanzees, bonobos, gorillas, and orangutans—it is worth asking ourselves, again, if our cousins on the evolution tree could be like us; if there’s a fundamental difference that separates us from them, or if that difference is just because they have been cast aside due to our dominance on Earth. We could even ask if in an alternate reality, they could surpass us as dramatized in *Planet of the Apes*.

In 1872, a little more than ten years after having presented his theory of evolution of species, Darwin published *The Expression of the Emotions in Man and Animals*. In this work he describes the notable similarity in the way different species express emotions (that is, mental states), reaffirming the notion of a biological affinity provided by evolution through common ancestors. Sometime later, in 1925, Wolfgang Köhler, one of the founders of the Gestalt theory in psychology, described the behavior of a quartet of chimpanzees called Chica, Grande, Konsul, and Sultan, who stacked boxes and joined sticks to reach bananas hanging beyond their reach. Interestingly, Köhler claimed that these chimpanzees didn't randomly find this solution, but instead they did so after taking some time for reflection and inspiration, after a few first failed attempts. In other words, their behavior responded to a process of reasoning and use of intelligence. Köhler's conclusions clearly contradict the up-to-then predominant ideas of *behaviorist* scientists, who believed animals only learn to associate stimuli with specific responses but without understanding what they are doing.



Sultan joining two sticks to reach an object and Grande reaching food using stacked boxes. (From *The Mentality of Apes*, Wolfgang Köhler, 1925.)

Köhler's description reminds me of a beautiful story told by Plutarch, and also discussed by Montaigne. A barber in ancient Rome had in his establishment a jay, a bird from the Corvidae family, who was a prodigy, imitating words or different sounds. One day, a group of musicians was playing outside the barbershop to the pleasure of passersby and, from that moment on, the jay did not emit a sound. It was thought that perhaps the sound of the bugles had deafened it and that such deficit in its hearing had affected its voice. But the reason was quite different. After listening to the music, the jay had plunged into a profound meditation, preparing its voice to imitate the new instruments and, after some time, it burst into song once again, repeating the melody and sound of the bugles and other instruments.[5]

Oddly enough, the jay has been a well-studied bird in neuroscience given its incredible memory ability. This bird stores acorns and seeds in the summer to survive the winter, but it must also hide food in different places, anticipating that some hiding spots might be discovered. What's surprising is that these birds remember the location of thousands of places distributed in a radius of several miles around their nests, and they're even capable of recognizing under the snow the exact spot where they left an acorn months earlier.[6] Furthermore, jays tend to steal food from each other, and if they notice another bird watching them when they store their food, they will come back later to change their hiding place and fool the eventual thief. This shows they have what is known as *theory of mind*, that they are capable of perceiving the point of view of someone else. In other words, the jay that stores food understands how the bird that is watching thinks and, by putting itself in the other's shoes, the jay can fool the would-be thief.

The ability to understand someone else's thoughts and emotions is far from trivial and is the foundation of our social interactions, communication, empathy, education, and deception. It is something that was also historically attributed exclusively to humans and that babies only acquire after a few years of life.

In a classic psychological test, a child subject watches a person hide a piece of candy in a box and then leave the room. In the adult's absence, another person moves the candy somewhere else. When the first adult returns to the room, the child has to indicate where they think this person will look for the candy. Only when children are a few years old do they tend to point to the first spot, showing they understand that what one person believes can differ from reality.

One of the most resounding scientific results of 2016 was an adaptation of this experiment using videos, and it clearly showed that chimpanzees, bonobos, and orangutans also possess theory of mind and can place themselves in someone else's shoes, understanding how others think according to what they believe, even if those beliefs are false. The researchers tracked the attentive eye movements of apes as they watched video scenarios designed to test their ability to predict where a person was going to look for an object that someone else covertly moved to another location.[7] Understanding that another animal has false beliefs is the basis of deception. Just like the jay changes the hiding spot of its acorns, chimpanzees tend to deceive each other, for example, diverting attention from an object of interest in the presence of other animals to later access it without competition, or distracting the dominant male in order to court a female.*[8]

In the same way we ask ourselves if animals can understand the thoughts of others, we can also ask ourselves if they are aware of their own existence. The two questions are actually quite related, because we tend to use the knowledge we have of ourselves to make predictions about the behavior of others. Obviously, we don't have a common language to ask animals about their consciousness in order to apply an evaluation similar to the Turing test described in chapter two. However, the American psychologist Gordon Gallup, Jr., found a very ingenious solution in the 1970s.

Realizing that chimpanzees seemed to recognize themselves in a mirror, Gallup designed the following test: While they were sleeping, he painted part of their eyebrows and ears with red dye, and when they woke, he placed them in front of a mirror. Gallup noticed that the animals began to regularly touch their painted parts, that is, they recognized themselves in the reflected image. As trivial as it may seem, very few animals can pass this test, among them higher primates, dolphins, and elephants.[9] Most animals believe that their reflection is another animal: chicks constantly chirp if they're alone but calm down when they're around other chicks or in front of a mirror; pigeons lay fewer eggs in isolation than if they're with other pigeons or in front of a mirror, and so on.

We've described undeniable intelligence traits in higher primates, but such achievements are not limited to manipulating objects to reach food (as demonstrated by

Köhler), using deceitful behavior, or recognizing themselves in a mirror. Jane Goodall, who for years studied chimpanzees in their natural habitat in Africa, also observed that they were capable of using tools—something that had been historically attributed exclusively to humans. For example, they used sticks placed inside anthills to acquire and eat termites, and they used rocks to crack nutshells. Sometime later, the husband and wife primatologist team Robert Seyfard and Dorothy Cheney observed that one monkey species in its natural habitat emitted different alarm cries according to the danger. Specifically, these animals had a distinct cry they used when spotting land predators (a leopard), another one for flying predators (an eagle), and yet another for slithering predators (a python). The reactions of the other monkeys differed according to the cry: they would climb a tree in response to the leopard cry, would look up and take shelter in the bushes when hearing the eagle alarm, or they would stand on two legs looking down to the ground after the python alert.[10]

A bonobo removing termites with a stick in the San Diego Zoo, and a capuchin monkey using a rock to crack a nutshell.*

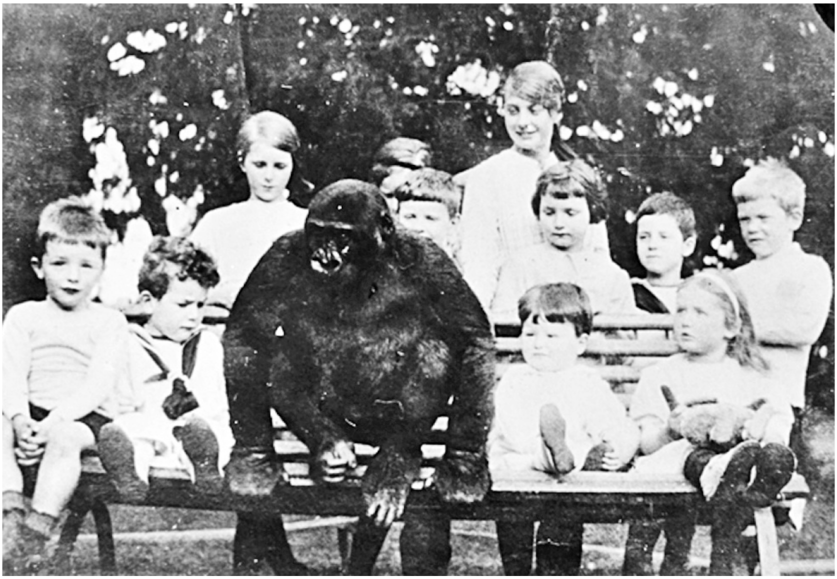
So, it is clear that intelligence is not just a human attribute and there is no fundamental difference, but rather different degrees of intelligence in other animals. Charles Darwin makes a masterful and detailed defense of this point:

> There can be no doubt that the difference between the mind of the lowest man and that of the highest animal is immense. An anthropomorphous ape, if he could take a dispassionate view of his own case, would admit that though he could form an artful plan to plunder a garden—though he could use stones for fighting or for breaking open nuts, yet that the thought of fashioning a stone into a tool was quite beyond his scope. Still less, as he would admit, could he follow out a train of metaphysical reasoning, or solve a mathematical problem, reflect on God, or admire a grand natural scene . . .
>
> Nevertheless the difference in mind between man and the higher animals, great as it is, is certainly one of degree and not kind.
>
> —Charles Darwin, *The Descent of Man*, 1871

Let us then ponder this question: How much of the difference between our intelligence and that of other most advanced animals is attributable to a cultural matter, a side effect of how we are educated, and how much is due to the process of genetic evolution?* Like the stories of Tarzan or Mowgli (from *The Jungle Book*), which show how the life of a child raised outside of civilization might be, we can similarly ask ourselves how intelligent could an ape be if it were raised and educated as a human.

Coincidentally, while I was writing about this subject, a story hit the British tabloids about “Mr. John Daniel,” a peculiar inhabitant of a small English country town in the early twentieth century. As a youngster, John Daniel attended school and played with his classmates, he made his bed, helped wash the dishes, and was even allowed to have tea with the adults. But John Daniel wasn’t just another boy; he was a gorilla adopted a few months after his birth and raised like a son by an affectionate and affluent woman. The incredible story of this gorilla, which particularly enjoyed drinking cider and was considered one more among the town residents, unfortunately does not have a happy ending.

When he became an adult and grew larger and heavier, it was impossible for his caretaker to continue keeping him in her house. So she decided to sell him to a company specializing in the animal trade business, with the promise that he would go to a park in Florida. However, John Daniel ended up in a circus and died shortly thereafter of pneumonia. His keepers back then attributed his death to a state of depression that led him to stop eating for long periods of time. John Daniel missed his home and his native village, going from growing up playing with other children to being a curiosity exhibited in a cage.



Gorilla John Daniel with students at a school in England around the 1920s (*Gloucestershire Live*/SWNS).

The story of John Daniel is in some ways related to the premise of one of the sagas of *Planet of the Apes*. The original movie’s commercial success led to four sequels, released between 1970 and 1973, a TV series, different comic books and a video game, a remake directed by Tim Burton in 2001, and a trilogy between 2011 and 2017. In the first part of the trilogy, *Rise of the Planet of the Apes*, Will Rodman, a scientist at a lab that tests Alzheimer's drugs on monkeys, discovers a newborn only a few days old when checking the cage of a chimpanzee that had just been sacrificed. Feeling sorry for the poor little animal, he decides to secretly take it home. The chimpanzee, named Caesar, lives in Rodman's attic and following the young man's patient daily lessons, with time begins to show noticeable progress in his learning.

The idea of raising an ape like a human has not only piqued the curiosity of sci-fi writers, but also of many researchers. In the 1940s, Keith and Catherine Hayes, married scientists who worked in a primate lab in Florida, realized that comparing the behavior of a human with that of an ape raised in the jungle, or in the limited confinement of a lab, didn't make much sense. For this reason, they decided to adopt a chimpanzee called Viki only days after she was born, and they raised her like another member of their family. They filmed Viki's progress as she matured, and they observed that in many aspects, the chimpanzee behaved like a child of the same age, learning based on imitation.

Viki played with different toys and objects, and with time learned how to dress herself, go to the bathroom, eat from a plate using a spoon, open cans and bottles, pour a cup of coffee, drink chocolate from a cup and add cold water if it was too hot, light matches, and even carry out complex motor tasks like using a hammer, a saw, a nail file, or a screwdriver. Replicating what Köhler's chimpanzees did in previous studies, Viki was also capable of joining sticks and stacking boxes to reach objects hanging from a tree. To have even more convincing proof of Viki's intelligence, the couple tried to teach her to speak. However, her progress in this aspect was much more limited, and after several years of learning, she was only able to say four words: "mom," "dad," "up,' and "cup."

Sometime later, in the 1960s, married psychologists Beatrix and Allen Gardner noticed that the shortcomings in ape speech learning was due to the structure of their vocal tract, which doesn't allow them to vocalize like humans. The Gardners found a brilliant way of solving this problem. Based at the University of Nevada, Reno, they cross-fostered a chimpanzee named Washoe, and instead of trying to get her to speak verbally, they showed her how to communicate using sign language. Washoe learned a few hundred signs, which to a certain extent she was also able to combine. For example, when she first saw a duck, she signed "water-bird," and another time she said a watermelon was a "candy fruit."

Not everyone agrees that Washoe or other apes could communicate using language, as humans do. For linguists like Noam Chomsky (popularly known for his positions in both science and politics), language is an innate and exclusive human ability. Diametrically opposed to what was argued with Washoe, Viki, and other experiments, Chomsky's view is in line with Descartes's thoughts and echoed by Steven Pinker, another famous linguist, who masterfully illustrates this point with the following analogy:

> People know how to talk in more or less the sense that spiders know how to spin webs. Web-spinning was not invented by some unsung spider genius and does not depend on having had the right education . . . spiders spin spiderwebs because they have spider brains . . .
>
> —Steven Pinker, *The Language Instinct*, 1994

The First Signs of Washoe, the 1974 PBS documentary that shows Washoe's feats, pointedly opens with quotes by Chomsky and other contemporary thinkers that the

Gardners were trying to prove wrong. However, the Gardners' conclusion wasn't universally accepted. After carefully analyzing videos of Washoe, Herbert Terrace, a psychologist from Columbia University in New York, criticized the results, claiming that the chimpanzee only repeated the signs that were made by its trainers, without showing the ability of language. In fact, language goes way beyond the attribution of signs to different concepts; it also involves the use of syntax to logically relate words and express thoughts.

Determined to surpass this criticism and thus demonstrate that Chomsky and his colleagues were wrong, Terrace planned a scientific project that would reveal the possibility of "humanizing" a monkey. The idea was to show that a chimpanzee raised and educated as a human could speak using sign language like Washoe. However, contrary to the approach taken with Washoe, Terrace's primary objective was to show that a chimpanzee could understand syntax rules, so as to form simple sentences and differentiate, for example, between "John gave the ball to Peter" and "Peter gave the ball to John." The ultimate goal was to communicate with an ape to exchange information and suss out what these animals think and feel.

The project ended up being a disaster, not only from a scientific point of view but from an ethical and human perspective. The story of the poor chimpanzee, called Nim Chimpsky in clear reference to Noam Chomsky, is portrayed in the 2011 British documentary *Project Nim* (based on Elizabeth Hess's book), which describes the negligence with which the animal was treated and the disregard and lack of ethics with which Terrace carried out his project.

Oddly enough, *Project Nim* and *Rise of the Planet of the Apes* were released at the same time, and their premises have remarkable coincidences—and not by chance. In an interview, one of the *Rise of the Planet of the Apes* screenwriters said he was inspired by Nim's story. I would hazard to guess that the documentary makers may have planned to release their work in tandem with the movie to take advantage of the widespread promotion it would give to their depiction of the real story.

Just as Caesar in the *Planet of the Apes* remake was adopted by a scientist, Nim was given to Stephanie LaFarge, one of Terrace's former doctoral students, who raised him at home together with the rest of her children. However, in that household, no one knew sign language to be able to teach Nim how to communicate effectively. After years of the chimpanzee living like one more member of the LaFarge family, Terrace chose to move Nim out and leave him in the care of another student. Then he passed Nim on to yet another student, and so devolved Nim's first years, having as many as three different tutors and dealing with various students who tried to teach him sign language.

In the remake of *Planet of the Apes*, Will Rodman, the scientist in charge of Caesar, decides he has no choice but to give him up for adoption to an animal shelter after an altercation with a neighbor. The scene where Rodman leaves an incredulous Caesar with other apes is perhaps the movie's highest dramatic point. (With time, Caesar plans a massive escape for the apes and takes them to shelter in the redwood forest north of San Francisco. These apes are the seed of the new civilization that will end up dominating Earth in the future. But that's another story.)

In real life, Nim never managed to organize any revolution but unfortunately ended up in an animal shelter like Caesar. After running out of funds, Terrace evaluated that he had enough information and unexpectedly decided to cancel the project. It is true that Nim became more aggressive as an adult,* and it is also true that it's hard to continue a scientific project with no funding. But these hurdles do not excuse the lack of any contingency plan nor the improvised and irresponsible behavior that characterized the project. One of the literary criticisms of the movie perfectly exposed the problem at length, quoting a dialogue between the Little Prince and the fox:

> "People have forgotten this truth," said the fox, "but you mustn't forget it. You are forever responsible for what you have tamed."
>
> —Antoine de Saint-Exupéry, *The Little Prince*, 1943

The testimonies of the last people in charge of Nim, describing the moment when they had to leave him at the animal shelter, are truly gut-wrenching. The sadness this produces is infinitely more profound than the equivalent scene in *Rise of the Planet of the Apes* because, in this case, the story is true. Nim would later be transferred to a lab where they would use him to test different vaccines and, in captivity, he would try unsuccessfully to communicate with his keepers and the other chimpanzees using the sign language he had been taught.

Sometime later, Stephanie LaFarge visited Nim, who by this time was retired in an animal sanctuary. He recognized her instantly, and his keepers observed that he became very nervous and potentially aggressive. Disregarding their opinion, LaFarge insisted on entering the cage. In that extremely dramatic moment, reuniting with his first mother after many years, Nim at first moved away, but then grabbed her by the ankle and started spinning her violently on the ground. However, he did not kill her. According to LaFarge, he wanted to show her that he was greatly disappointed with how she had abandoned him to such a miserable and unfair fate.

I personally confess that watching *Project Nim*, and writing about this subject, is painful, but I believe absolutely necessary to remind us of the obligation we have, as a highest species, with other animals. This is particularly important for scientists. I support animal research and condemn violent acts as well as biased opinions of activists who dogmatically oppose everything and try to instill fear in the scientific community. I believe these experiments are necessary, but they must be very carefully regulated by ethics committees that guarantee there are no repeats of the Nim fiasco. I once heard the Dalai Lama—the spiritual leader of Buddhists, who believe all animals have a certain degree of consciousness and, therefore, avoid even killing a mosquito—say that he thought animal testing was necessary, but they had to be treated with compassion.* His words perfectly summarize my point of view on this matter.

Beyond the ethical aspects, Terrace's project was a complete scientific failure given its serious flaws. Foremost, Nim was trained by several different people in structured learning sessions, which is like learning a second language at school and very different from learning how to speak through natural daily interaction with a human family. The project ended prematurely while Nim was still learning, and the instability in his closest environment, changing his caretaker several times, certainly did not favor his education. In the end, Nim wasn't able to do more than Washoe decades earlier. Terrace ended up admitting his failure, agreeing with the group of linguists led by Chomsky, who claimed that animals weren't capable of developing language and that what appeared to be communication was nothing more than an imitation of the gestures of their tutors to make different requests: play, eat, drink water. Yet Chomsky's conclusion would yet again be challenged by another extraordinary animal.

The bonobos, native to the Congo, are very similar to chimpanzees, although somewhat shorter, with a much less aggressive character, a straighter posture, and a generally twolegged walk. A bonobo's build is, in fact, very similar to the famous australopithecine dubbed Lucy, whose partial skeleton, dating back three and a half million years, was discovered in Ethiopia. Kanzi, a bonobo born in captivity, would play and get into mischief while Sue Savage-Rumbaugh (one of the few scientists who continued studying the ability of language in primates after Terrace's failure) fruitlessly tried to teach Matata, his adoptive mother, to communicate using an electronic board with different symbols. One day Matata was temporarily transferred to another center, and Kanzi felt the irrepressible need to use the board to ask about his mom. What's surprising is that Kanzi had never had a training session and just learned how to use the board by watching how Savage-Rumbaugh taught his mother. The big difference with Nim or Washoe is that Kanzi didn't use signs to ask for something specific; he needed to communicate. After this episode, Savage-Rumbaugh realized that beyond teaching them words, she had to stimulate the bonobos' need to "talk," sharing games, food, long excursions, and camping, among many other things.

With time, Kanzi learned hundreds of words and began combining them and understanding the meaning of complex sentences, even those referring to past events or involving objects and places that weren't in plain sight. One of the biggest limitations of not having a language is that animals are only able to exchange information about the immediate present and not about past experiences or future plans. Kanzi also demonstrated some understanding of the syntactic meaning of sentences, that is, recognizing that language goes beyond a mere association of words, and that their order and structure represent different meanings. For example, when in a video Savage-Rumbaugh asked another bonobo called Tamule to hug Kanzi, Kanzi didn't hug Tamule but rather grabbed Tamule's arm and put it around him. Kanzi allegedly recognized the difference between "Tamule, hug Kanzi" and "Kanzi, hug Tamule." Among many other things, Kanzi could also cut branches and light a fire on an excursion, and even play *Pac-Man*. Savage-Rumbaugh says that Kanzi also enjoys watching movies and one of his favorites is actually *Planet of the Apes*.

Panashiba, Matata's biological daughter and Kanzi's half-sister, also demonstrated the ability to communicate, and her reputation led her to even improvise on a keyboard with rock musician Peter Gabriel of Genesis fame. But language ability is not restricted to bonobos and was likewise shown by other higher primates, such as the orangutan Chantek and the gorilla Koko, who was Michael Crichton's inspiration when writing the sci-fi novel *Congo* (later adapted to the big screen with the same title).

In these last pages, we focused almost exclusively on language, since our use of language appears to distinguish us most from other species. However, we could argue that it may be an error to focus on language as an almost exclusive sign of intelligence, because when testing the language ability in different animals, we may be making an anthropomorphizing mistake, evaluating them by our standards. Sextus Empiricus claimed that animals have different mental images due to great differences in their senses:

> Perfume appears very pleasant to humans but intolerable to dung beetles and bees. Olive oil is used by humans but destroys wasps and bees if poured over them . . . If the same things for some are unpleasant and for others are pleasant, and if the pleasant and unpleasant resides in the imagination, then animals have different mental representations of objects.
>
> —Sextus Empiricus, *Outlines of Pyrrhonism* (XIV: 55, 58)

Much like Descartes, we tend to assume that, similar to humans, any intelligent animal would want to express its thoughts. But the fact that animals don't speak among themselves could be because they have different cognitive representations and, by feeling the world differently, perhaps they don't have the need to communicate, or maybe they even do it through other means. Austrian philosopher Ludwig Wittgenstein (whom we will discuss further in the following chapter) said in this regard that "if the lion could speak, we wouldn't be able to understand it." Besides these exceptions, the use of language leads to two big benefits: first, it allows us to communicate and transmit knowledge and, secondly, it reinforces the ability of abstraction, of classifying and constructing concepts. Language lets us go beyond thinking exclusively about things within our line of sight in our immediate present; it allows us to label concepts and think about our thoughts, assess the future, and meditate on the past.* More specifically, words are abstractions of reality. When I say "dog," I'm not referring to my childhood dog, or to the one of my neighbor; it doesn't matter if it is big or small, a guard, a hunter, or if it has a light-colored coat with spots on its torso. When I say "dog," I set all of these details aside and am referring to what defines the animal. In a fantastic essay, Jorge Luis Borges, perhaps the most influential Argentine writer of the twentieth century, says:

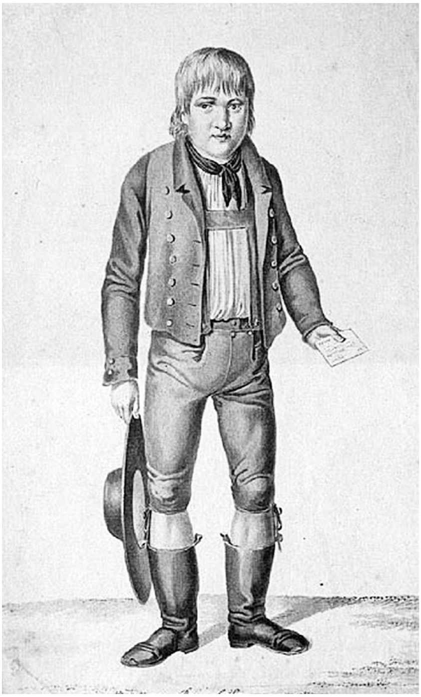
> The apparent world is a pile of shuffled feelings . . . Language is an effective management of that enigmatic abundance of the world. In other words: We invented nouns for reality. We sense a circle, see a small heap of light the color of dawn, a tickle brightens our mouth, and we lie, saying those three heterogeneous things are one and that it is called orange. The moon itself is fiction. Aside from astronomic conveniences that shouldn't busy us here, there is absolutely no similarity between the yellow circle that is now clearly rising over the Recoleta wall and the sliver of pink I saw in the Plaza de Mayo sky several nights ago. Every noun is an abbreviation. Instead of saying cold, sharp, hurtful, unbreakable, shiny, pointed, we state dagger; instead of the distancing sun and shadow profession, we say dusk.
>
> —Jorge Luis Borges, *El Tamaño De Mi Esperanza*, 1926

Furthermore, in the short story "Tlön, Uqbar, Orbis Tertius," Borges conceives a world where language has no nouns and concludes that in this place, "Every mental state is unyielding: the mere fact of naming it—*id est*, classifying it—imports falseness. From there, it can be concluded that there is no science in Tlön—not even reasoning."*

In the late seventeenth century, John Locke (1632–1704), one of the most notable exponents of British empiricism, famously opined, "Brutes abstract not."[11] In his often quoted *An Essay Concerning Human Understanding*, Locke based this statement on the fact that animals don't use words or some type of sign to represent general ideas. To some extent, nowadays we could reject Locke's affirmation with cases like Kanzi, Koko, or Chantek. Animals, and particularly higher primates, are much more intelligent than Aristotle and Descartes believed, and they have at least a limited language ability. However, none of these animals writes poetry, asks itself about the origin of the universe, or solves integers. This leaves us wondering if the distinction between humans and apes is only a question of environment and language use, or if there are more fundamental differences.

In the spring of 1828, a teenager was found wandering the streets of Nuremberg, Germany. The boy, around sixteen years old, could barely speak or demonstrate awareness of himself. He said his name was Kaspar Hauser and claimed to have lived all his life confined in a dark cell, with no other human contact. Kaspar's story created a rumor mill and had all of Germany on tenterhooks for a long time. It was even believed that he was the crown prince of Baden, which was much later disproved by a DNA analysis.

The true story continues to be a mystery worthy of a Brothers Grimm plot. There are also some who say that Edgar Rice Burroughs based the Tarzan character on a true story, and many other tales exist of men who were raised outside of civilization, from the legend of twin brothers Romulus and Remus (the mythical founders of Rome) raised by a she-wolf, to the most recent case of a Vietnamese father who took his newborn boy to live in the jungle, escaping the war. Beyond how much we want to believe these stories or not, it is clear that a human raised in solitude, completely removed from civilization, couldn't reach the intellectual development of a typical person. However, such a person could quickly learn endless tasks that would be impossible for a bonobo such as Kanzi, even though the latter was raised around humans.



Portrait of Kaspar Hauser.

The feats by primates like Kanzi are impressive, but they are way behind those shown by a two-year-old child. I don't deny the intelligence of animals at all, but there is a clear difference from the abilities of humans. This difference is not just educational but also genetic. It probably emerged after many years of evolution when one species, humans, developed the ability of deep and abstract thought facilitated by the use of language.

In chapter two, I described the discovery of concept cells in a key memory area in the human brain. One of these neurons responded to Jennifer Aniston, another one fired up when showing a subject different pictures of Halle Berry, and another did so with Oprah Winfrey. What's important is that these neurons responded selectively to pictures of specific people and even these people's written or spoken names; in other words, they don't code details but rather concepts. And it makes sense that there are neurons representing concepts in an area involved in memory, because we tend to store the meaning of things and forget details. This is the difference that we underscored when comparing ourselves with computers, and I believe it is also a key difference with other animals. What's interesting is that no one has yet found neurons like these in other species. It is possible that in a few years someone may report a similar coding of concepts in monkeys, but for now, and after several attempts in different labs, this isn't the case.

But do we have language because we have an explicit coding of concepts in the brain, or do we have an explicit coding of concepts because we have language? I dare speculate that both things evolved simultaneously, reinforcing each other. Let me clarify that this argument is very speculative; it is just a personal opinion, although I refuse to think that the fact that we are the only species with language and concept cells (at least for now) is a coincidence.

Perhaps what most resembles concept cells are *place cells*—neurons that represent specific places—in rodents. If changes are made in the environment where the rodent carries out an experiment, these neurons show what is known as *remapping*, that is, they begin to respond in different places and, typically, when the environment changes, the rodent that learned how to carry out a complex task has to relearn it practically from scratch.[12] In other words, the rodents aren't capable of generalizing and realizing that they are doing the exact same thing but in a different situation.

Usually, rodents tend to learn tasks independently, without looking for a connection among them or shortcuts to facilitate their learning. The ability to generalize, to apply knowledge acquired in a different situation or circumstance, is precisely what we prevail at as humans. As noted in previous chapters, a similar difference exists between the development of machines and algorithms for specific tasks, and how elusive it has been to create circuits capable of making abstractions and generalizations, showing general intelligence. But such intelligence comes at a price. Of all the species, we take the longest to mature.[*] The difference is that from the start, the offspring of other animals learn specific tasks to survive. Meanwhile, a human baby has a much slower development process precisely because it is establishing the foundation of its intelligence. The baby cannot deal with basic needs on its own, but the adult that the child becomes will one day wonder about their place in the universe and what differentiates them from other species.

---

[*] We will see that this is actually not true. Specifically, the layout of the vocal tract in monkeys does not allow them to vocalize like humans.

[**] Nowadays, we know that this argument by Descartes is also not correct. Specifically, a parrot named Alex, studied by scientist Irene Pepperberg, demonstrates understanding in what it says. For example, when Pepperberg asks what color a certain object is, Alex doesn't mechanically repeat the questions, like other parrots, but instead gives the right answer. According to Pepperberg, among other things, Alex can recognize around fifty objects, different shapes and colors, concepts like *big, small, equal*, and *different*, and quantities up to six.

[*] The idea of putting yourself in someone else's shoes to deceive them is the quintessence of magic. A magician's ruse consists precisely of implanting false beliefs in the spectator that allow for unexpected resolutions. For more details, see, for example: Rodrigo Quian Quiroga, "Magic and Cognitive Neuroscience," *Current Biology* 26, no. 10 (2016): R387–R407.

* Image of capuchin monkey reprinted from Dorothy M. Fragaszy et al., “Wild Bearded Capuchin Monkeys (*Sapajus libidinosus*) Strategically Place Nuts in a Stable Position During Nut-Cracking,” *PLOS One* 8, no 2 (2013): e56182.

* This is part of a much more general and renowned discussion known as *nature versus nurture*.

* As described by Jane Goodall, the great difficulty of raising chimpanzees in human homes is that they tend to become aggressive as they mature.

* I remember the Dalai Lama saying this in response to an audience member’s question, after a talk he gave at the Society for Neuroscience’s annual meeting in 2005.

* Philosopher Daniel C. Dennett has a brilliant description of these ideas in chapter five of his book *Kinds of Minds* (New York: Basic Books, 1996).

* “Tlön, Uqbar, Orbis Tertius” was first published in the Argentine journal *Sur* in 1940 and later in the collection *Ficciones* (*Fictions*) in 1944. For a more comprehensive description on the thoughts of Borges and other philosophers about abstraction and language, see chapters two and twelve of *Borges and Memory* (cited in chapter one).

* This should be also attributed to the fact that babies are born with only about one third of their brain developed due to what is called the *obstetric dilemma*; this means if we were born with a fully developed brain, the head would be way too big and labor would be impossible. Nature has found this solution to our relatively huge brain size. For more details see Robin Dunbar, *Human Evolution*, (London: Penguin, 2014), 245.

CHAPTER 4

THE MATRIX

The Illusion of Reality

A young, dark-haired man is slumped in front of his computer, asleep while music plays through his earphones. His desk is overflowing with papers, keyboards, cables, printers, data disks, and other gadgets, and the monitor shows sequences of automatic online searches, typical of a hacker. Suddenly, the screen goes dark. In the dimly lit room, the man opens his eyes and sees, line by line, a cryptic message in green fluorescent characters, like those used in old-school computer monitors:

> Wake up, Neo . . .
> The Matrix has you . . .
> Follow the white rabbit.
>
> —*The Matrix*

Half asleep, Neo tries different commands to clear the message from the screen, and a new line appears: "Knock, knock, Neo." On cue, there's a knock at his door. A guy has come to pick up a disk we assume contains pirated data.* He has a small entourage with him, including a woman with a white rabbit inked on her shoulder. She encourages Neo to join their evening excursion. Intrigued, he agrees to go with them to a nightclub. There he meets Trinity, the elusive hacker responsible for the message on his computer and who eventually takes him to meet Morpheus, the leader of a secret gang of partisans. Neo is obsessed with one question: What is the Matrix? He believes that Morpheus has the answer.

Morpheus holds an audience with Neo in an abandoned hotel room. As they sit face-to-face in two imposing leather chairs, Morpheus begins to speak of the Matrix, but his statements are perplexingly vague:

> **Morpheus:** The Matrix is everywhere. It is all around us. Even now, in this very room . . . It is the world that has been pulled over your eyes to blind you from the truth . . . That you are a slave, Neo. Like everyone else, you were born into bondage. Born into a prison that you cannot smell, or taste, or touch. A prison for your mind.
>
> —*The Matrix*

Morpheus opens his hands and shows Neo two small pills. If Neo takes the blue pill, explains Morpheus, the story will end, and he will wake up in his bed thinking it was all a dream. However, if he takes the red pill, he will access the truth and learn for himself what the Matrix is.

As we expect, Neo takes the red pill and is rescued from the illusion that was his life up until then—only to face a devastating reality: The Matrix is a construction created by a self-serving artificial intelligence to appease the minds of unsuspecting humans, who are maintained from birth in incubator pods so that energy can be generated through their nerve impulses. After a cruel awakening from the "dream world" imposed by the Matrix, Neo must recover from extreme muscular atrophy, having never actually opened his eyes or used his body. And once he is fit, Morpheus shows Neo how the Matrix operates, using computer simulations plugged directly into his mind. As he encounters these digitally constructed realities, Neo argues that they are not real, to which Morpheus replies:

> **Morpheus:** What is real? How do you define real? If you're talking about what you can feel, what you can smell, what you can taste and see, then real is simply electrical signals interpreted by your brain.

Morpheus grabs a TV remote control, and while different images of the daily reality Neo has known flash across an oldfangled television screen, he continues:

> This is the world that you know. The world as it was at the end of the twentieth century. It exists now only as part of a neural-interactive simulation that we call the Matrix.
>
> —*The Matrix*

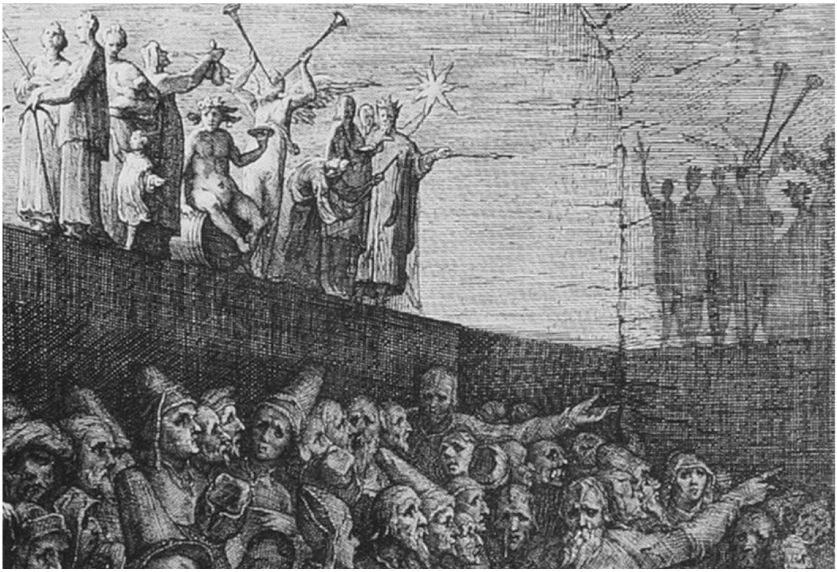
The idea behind *The Matrix* is not so original. In fact, the notion of an alternative reality in cyberspace is the basis for works such as *Neuromancer*, by William Gibson, one of the great pillars of the cyberpunk movement. This trend within sci-fi presents the contrast between the enormous advances of technology (and the use of cyberspace) and the miserable reality of marginalized individuals, typically hackers, who defy the societal system and the big corporations. The cyberpunk genre of stories and movies was born from a reaction to classic sci-fi works, which tend to describe epic interstellar journeys and future civilizations. In contrast, cyberpunk reflects on the impact of technology and on what it means to be human.

Aside from its recycled concept, *The Matrix* was also criticized for its excessive adrenaline, including what were perhaps exploitative kung fu combat scenes and marathon confrontations with automatic weaponry.

In neuroscience, there's a type of memory called *flashbulb*. These are memories that resonated so deeply, positively or negatively, that years later we can bring them to mind with incredible richness of detail. For an Argentine about my age, a flashbulb memory is, for example, Maradona's goal against England in 1986, or the 1978 World Cup final against Holland. I have argued more than once that, at least for the typical Argentine, most flashbulb memories are connected to soccer games. So, to illustrate how much I liked *The Matrix*, it suffices to say that I remember exactly the theater where I saw the movie a couple decades ago, and how I ended up wandering the streets of Aachen (back then I was doing a postdoctoral degree in Germany), elated by the story I had just seen and the closing music by Rage Against the Machine replaying over and over again in my head. I couldn't have cared less about the plot's unoriginality or excessive action. I enjoyed it like crazy.

Yet the underlying idea in *The Matrix* dates back not only to the cyberpunk movement but much earlier in time. Precisely one of the most transcendental questions in philosophy is related to the existence of the self and the external world. It is as simple and yet as profound as asking oneself, "Do I exist? Does what I see around me exist?"

In *The Republic: Book VII*, written around 380 BC, Plato presents his famous allegory of the cave: A group of people live chained inside a cave and can see only the shadows of what happens behind them. For these people, the shadows are their only reality, since they cannot see what generates them. In this way, Plato argues that we see only manifestations of reality, since reality itself is beyond our perception. For example, if we see one or another horse, we perceive specific manifestations from which we abstract the universal concept *horse*, which has an existence independent from the various manifestations and even from ourselves.



Detail of the print created from Cornelis van Haarlem's *Antrum Platonicum* (Jan Saenredam, 1604). In Plato's allegory, the cave people (*at the bottom of the piece*) can see only the shadows projected on the wall (*upper right*) and not what produces them (*upper left*).

This philosophical conception, known as *platonic realism*, states that there is a universal reality. For Plato, a philosopher's job is to make use of reason to access external reality (the concept of "horse") based on the different manifestations that reach him through the senses (the different horses he observes).

Whereas Plato compared sensory impressions with the shadows seen by people imprisoned in a cave, two thousand years later, Descartes would push to the limit the idea of doubting reality. In his quest to find the ultimate truths, those that are irrefutable, Descartes systematically doubted everything. To that end, in the first of his *Meditations* he conceived the possibility of an evil and powerful genius who uses all of his resources to deceive us regarding the external world. Although we may consider this very unlikely, it is an ingenious and valid philosophical argument because, in principle, we cannot reject the possibility that the "sky, air, earth, colors, figures, sounds, and all other external things are nothing more than illusions and dreams used by [the evil genius] to trap [our] credulity."[1] However, Descartes further argued that regardless of how sophisticated the deception may be, we can still confirm our own existence because we are capable of raising this question. Our ability to think, according to Descartes, is what ensures that we exist, and based on this idea, he states the most noted phrase in the history of philosophy: "*Cogito ergo sum*," which translates to “I think, therefore I am.”

With the previous arguments, Descartes establishes the existence of himself. In the subsequent *Meditations*, he proceeds to also demonstrate the existence of the external world, arguing that God is not fallacious and doesn’t tend to deceive us because that would make him imperfect. The existence of both thought and the external world would eventually lead Descartes to separate mind from matter, which, as we discussed in the first chapter, is the fundamental tenet of Cartesian dualism.

The opposing position to dualism is *monism*, that is, stating the existence of only the material world (*materialism*) or the world of ideas (*ontological idealism*). In ancient Greece, Democritus (460–370 BC), the forefather of atomic theory and one of the first exponents of materialism, claimed that there is nothing more than atoms and the space between them. The modern view of materialism is to assume that the mind is nothing more than the activity of neurons, as discussed in detail when we covered *2001*’s HAL 9000 and the machine’s ability to think. Ontological idealism claims the exact opposite: The only things that exist are ideas and the mind. Its most fervent supporter was Bishop George Berkeley (1685–1753), one of the most notable exponents (together with John Locke and David Hume) of British empiricism. For these philosophers, the mind is like a *tabula rasa*, or a blank slate, on which we record knowledge through experience and perception with our five senses.

Just like Descartes’s main premise was *cogito ergo sum*, Berkeley’s would be *esse est percipi*: to be is to be perceived. Berkeley argued that the external reality was nothing more than a group of ideas and contended that only our perception of things exists and not the things by themselves. For Berkeley, what is not perceived does not exist. But, of course, there are also minds that perceive and think.*

Later, the German philosopher Immanuel Kant (1724–1804) would propose a less radical position that we now know as *transcendental idealism*. Kant argued that we only get to know our representation of things, but never *die Dinge an sich* (the things in and of themselves). Such representations are created based on an innate intuition of space and time, meaning we conceive things in a time and a place. However, contrary to Berkeley, Kant did not deny the existence of things (and, therefore, his position does not oppose dualism); he only argued that we cannot get to know them by themselves, independently from the perception we have of them.[2]

Descartes used the evil genius argument to test (and later demonstrate) his own existence, and by believing in a benevolent God who doesn’t mean to hurt us, he also reaffirms the existence of the external world. However, if we set aside the conclusions based on the existence of a benevolent god, the evil genius argument leads us straight to the scenario of *The Matrix*, or what in philosophy is known as *solipsism*.** However, it is worth noting a difference. While the premise of *The Matrix* stipulates that countless people live in a fictitious reality implanted in their minds, solipsism is idealism taken to the extreme: not only does the external world not exist, but no mind other than ours exists. It seems insane, but how can we reject it? How can we prove that the people we see around us actually exist?

We shouldn’t be surprised that there is no record of a solipsist philosopher. Beyond the shortcomings we will get into later on, to spread solipsism would be a contradiction. If a person really believes they are the only one in the universe, what purpose does it serve to spread their ideas and try to convince those who are no more than an illusion of their senses? As Bertrand Russell said:

> As against solipsism it is to be said, in the first place, that it is psychologically impossible to believe, and is rejected in fact even by those who mean to accept it. I once received a letter from an eminent logician, Mrs. Christine Ladd-Franklin, saying that she was a solipsist, and was surprised that there were no others. Coming from a logician and a solipsist, her surprise surprised me.
>
> —Bertrand Russell, *Human Knowledge: Its Scope and Limits*, 1948

The modern version of Descartes's evil genius and solipsism is what is known as the *brain in a vat* argument drafted by American philosopher Hilary Putnam in 1981.* Let's imagine an evil scientist (or Descartes's genius), who surgically extracts our brain and places it in a big jar with a solution overflowing with nutrients to keep it alive. Let's also imagine that he connects all of our nerve terminals to an incredibly powerful supercomputer, capable of simulating the external reality by using complex electrical stimulations of different groups of neurons. For example, the supercomputer stimulates neurons that represent the concepts of sand, palm trees, and the sea, and leads us to believe we are on a beach. The argument, similar to the construct of *The Matrix*, is somewhat crazy but very interesting, because if we lack arguments to demonstrate that we are not a brain in a vat, then we can't state anything about the existence of the external world.

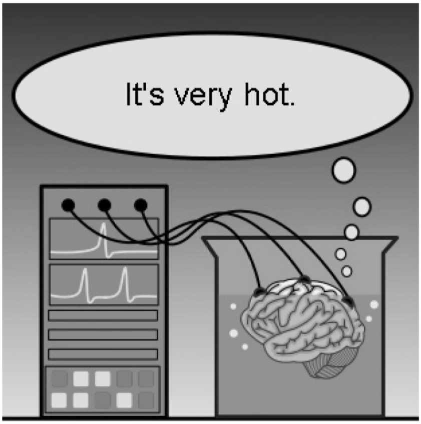

It's very hot.

Illustration of the brain in a vat argument. Through stimuli provided by a supercomputer, the brain believes, for example, that it is on a beach.

In neuroscience, there are several examples of how directly stimulating the brain can alter and even create different feelings. For example, when electrically stimulating in monkeys an area involved in the perception of movement (which in neuroscience is known as *MT* or *V5*), scientists from Stanford University showed that this stimulation changed the animals' perception of the direction of motion of a group of dots on a screen.[3] Likewise, a group of Iranian researchers showed that electrical stimulation of an area dedicated to facial recognition (the inferotemporal cortex) made the monkeys tend to report that blurry images they were shown corresponded to faces and not to other objects.[4] But the most amazing studies describing the effects of electrical stimulation on the brain were done more than half a century ago by neurosurgeon Wilder Penfield with humans.[5]

In chapter two, we mentioned that epileptic patients who do not respond to medication could have surgery to remove the area that causes the epileptic seizures. The success of such surgeries depends, on one hand, on finding with the best possible precision the seizure focus (and as we discussed earlier, to do so, in many cases, intracranial electrodes are implanted in the brain) and, on the other hand, on not harming areas that are vital for the patient's functionality. For example, if by mistake part of the motor cortex is compromised during the surgery, the patient would suffer paralysis; if the visual cortex is compromised, the patient could become blind; and so on. Given that the location of these vital areas varies from patient to patient, Penfield developed an electric stimulation protocol to define their precise location. This procedure was done while the patients were awake and their brains were uncovered (accomplished by temporarily removing the upper part of the cranium), all the while using local anesthesia since the brain itself doesn't have pain receptors. Penfield then stimulated the areas around the sensory and motor cortices, defining where the stimulation produced a tickling sensation or movements of different parts of the body, like the fingers, the tongue, or the lips. He also stimulated around the language areas while asking the patient to name different objects, or he stimulated areas in the visual cortex and produced different hallucinations—patients said they saw colored stars, and so on.

The most spectacular results took place when Penfield stimulated the areas related to memory in the temporal lobe, because in many cases he was able to reactivate different memories of the patients, who seemed to be reliving their past like a movie. One patient claimed to remember being in her kitchen listening to the voice of her son who was playing outside, and said she was conscious of the neighborhood's sounds, the passing cars, and how they were a danger to her son. When stimulating a specific point in another patient's temporal cortex, she reported hearing the voice of a woman and a man who were calling her somewhere along a river, and the stimulation of another spot brought up a familiar memory of being in an office with desks and a man leaning on one of them with a pencil in his hand. Penfield contended that these memories were stored in the temporal lobe and that the stimulation simply evoked them, triggering the flow of consciousness, just as the taste of Marcel Proust's famous madeleine fired memories of his childhood in *Remembrance of Things Past*.*

I would also like to mention a very interesting result from my colleague Itzhak Fried, the neurosurgeon whom I have been working with for years on the study of neural recordings in epileptic patients—who is, incidentally, also a deep admirer of Penfield. Itzhak showed that stimulation of the supplementary motor area in a patient triggered his laughter. When they asked him (during the surgery) what he was laughing about, the patient said that the team of surgeons "looked very funny, all around him."[6]

To sum up, the results obtained by different groups of neuroscientists show that electrical stimulation can elicit different feelings and even evoke memories. So, neuroscience seems to support the underlying idea in *The Matrix*. However, we will soon see that this isn't really the case.

The Invention of Morel, first published in Argentina in 1940, is without a doubt Adolfo Bioy Casares's most accomplished work. Jorge Luis Borges (a close friend of Bioy's) said in his prologue that he doesn't think it is hyperbole to describe its plot as perfect. A fugitive arrives on an abandoned island where there is only a museum, a chapel, and a swimming pool. One day, he is awakened by the music of a phonograph and sees several couples in old-fashioned clothing dancing on the terrace facing the museum. The fugitive hides, fearing he will be discovered by the interlopers, among them a woman who leaves her companions behind to see the sunset over the rocks. After days of secretly watching her, he falls hopelessly in love. Finally, he decides to talk to her, but quite to his dismay she doesn't even bother to answer him; she continues gazing at the sunset and reading her book, as if he didn't exist. Likewise, in different situations, the fugitive believes he's been discovered by the intruders and waits for them to find him on the beach. However, they seem to completely dismiss his presence, even after he finds himself unexpectedly standing face-to-face with one of them.

The story's resolution is astonishing: The dancing couples are no more than projections, with incredible realism, of something that was experienced by a group of friends on the island several years earlier. Why did I stop to tell this story? Because the details of the plot, so praised by Borges, hold the solution to *The Matrix* conundrum and the conception of the brain in a vat.

For the premise of *The Invention of Morel* to work, it is vital that the fugitive does not interact with the dancing couples and that this lack of interaction occurs naturally. Bioy's justification is perfect: The fugitive, escaping the law, does not want to be discovered, and the arrogant interlopers don't even deign to recognize his presence. If the woman watching the sunset had replied, or if that other person who was suddenly face-to-face with him had screamed in panic, the argument would fall apart. The reason is very simple: A group of unreal people cannot react or interact with someone of flesh and bone. Let's imagine for a moment that the sunset woman replied to the fugitive's words. What should she say? How would the brilliant mind that created these incredible projections, be it Morel, Descartes's evil genius, or the architect of the Matrix, anticipate the fugitive's question to give an appropriate response in the right moment? How could the wizard behind the curtain foresee if the reaction should come from an insult, a caress, or a declaration of love?

In the first pages of his masterpiece, *Consciousness Explained*, Daniel Dennett offers a brilliant description of this point. An imaginary reality could work as long as the person being deceived by that reality has no interactions. Once an exchange of information has been established, the deception becomes unsustainable.

A skilled illusionist could, for example, convince us that he is speaking onstage even though his image is nothing more than the projection of a sophisticated hologram. To solve the trick, he could suddenly make the image disappear and—abracadabra—appear amid the audience in one of the balconies. Of course, for the trick to work, no one must interrupt him and, furthermore, no one must get up onstage. The more freedom of movement he gives his audience, the more he interacts with them before disappearing—for example, pretending to have a casual conversation with a conspiring audience member—the more credible his projected image, and the more sophisticated (and spectacular) the deception will be, perfectly coordinated with such interactions.

Let's go back to *The Matrix*. How do I know the external world is real? Because it coherently responds to my expectations. Descartes's evil genius could, in principle, have my brain in a vat, making me believe that I'm on an idyllic beach. To do so, he would have to stimulate specific neurons in my visual cortex and induce in me the feeling of seeing the blue sea, the sand, the palm trees, and so on. When I touch the sand, he would also have to simulate in my somatosensory cortex the feeling generated by the sand sliding between my fingers; when I wade into the sea, he would have to simulate the change in temperature and the feeling of the water against my body; and if I splash around in the waves, each hand motion should produce tactile feelings and an incredibly intricate visual illusion of foam and thousands of drops in the wind. In other words, it may be possible to keep the illusion going while I lie flat on the sand, like in a nightmare where I can't move—just like it is possible for the illusionist to create the impression that a fictitious image is real, as long as he keeps his distance from the audience—but the moment I interact with the illusion, it becomes technically impossible to sustain, even for a powerful and evil genius. Dennett says that some philosophers argue that regardless of how difficult it may be to create such an illusion, in principle, it is possible. To which he answers, to my understanding, correctly, that it is also "in principle, possible to build a stainless-steel ladder to the moon or write, in alphabetical order, every possible conversation in English with less than one thousand words."[7]

The same argument is applied to the existence of others. It is impossible to simulate a coherent interaction with someone who doesn't exist, that is, specifying how and when this imaginary person should react to each of my actions. The variety of possible actions exponentially diverges as the interaction takes place, since even when talking about the same subject, the most frugal conversation can take endless paths.

The coup de grâce to solipsism and the possibility of a Matrix was delivered by Ludwig Wittgenstein (1889–1951), considered by many as the most brilliant exponent of philosophy in the twentieth century. Wittgenstein's story is so peculiar that it is worth briefly relating here. Born to a wealthy family in Austria, Wittgenstein studied aerospace engineering in Berlin and later in Manchester. His engineering studies sparked an interest in mathematics and its philosophical foundation, and eventually he became Bertrand Russell's disciple in Cambridge. Such was the passion he developed for philosophy that he often secluded himself for long periods of time to think about philosophical problems in an isolated cabin that he built himself on a Norwegian fjord.

When World War I broke out in 1914, Wittgenstein readied himself as an Austrian artillery volunteer, and he served and was decorated for his courage. He was captured as a prisoner of war toward the end of the conflict, and during wartime he finished the only book he would publish during his lifetime, the *Tractatus Logico-Philosophicus*. In the *Tractatus*, he considered the nature of the representations we make of things and concluded that such representations are given by thoughts, which in turn are like images. According to Wittgenstein, when we communicate, we exchange images. For example, when I tell someone I'm going on vacation to the Caribbean, I create in that person an image of myself lying on a beach. Wittgenstein concluded that many communication issues are due to misunderstandings created by the erroneous interpretation of the images we generate in the minds of others. Likewise, he argues that philosophy does nothing more than entangle things, causing endless misunderstanding, and in the last sentence of his book, he masterfully says in this regard: "Whereof one cannot speak, thereof one must be silent."

When he returned from war, Wittgenstein inherited his deceased father's fortune and became one of the richest people in Europe. However, he gave away his money and, believing he had solved the most fundamental problems of philosophy, he chose to work as a rural teacher and a gardener, and even took on the role of architect in the construction of his sister's home. A decade or so later, he began to doubt his principles and resumed his pursuit of philosophy in Cambridge. The result of this second stage is a book that was published postmortem, *Philosophical Investigations*, in which he vividly criticized his previous ideas.

In the *Tractatus*, Wittgenstein views words as representations: objective and absolute entities. In this posthumous work, he radically changes his point of view, arguing that words are given meaning by their context and use in the language. Such variability gives any clause a relative value, hinging on the subjective interpretation attributed to the clause's component words via their context. Therefore, there are no absolute criteria we can use to confirm the veracity or falseness of a clause; that judgment derives from our interaction with other people. In other words, through language we agree that some clauses are true and others are false. This is how Wittgenstein rejected the possibility of a *private language*; for example, we can state that the setting sun is red because we all agreed to call it that.* Wittgenstein says:

> Let's imagine this case. I want to keep a journal on the repetition of a specific feeling. To this end, I associate it with the symbol "S" and in a calendar I write this sign on every day I have the feeling . . . A definition works, incidentally, to establish the meaning of a symbol. Well, this happens precisely when focusing my attention, since, by doing so, I imprint the connection between the sign and the feeling. "I imprint it," however, can only mean: This process makes me remember in the future the correct connection. But in our case, I have no criteria about correction. One would say here: Anything I think in any case is correct, is correct. And this just means that here we cannot speak of "correct."
>
> —Ludwig Wittgenstein, *Philosophical Investigations*, sec. 258

Basically, Wittgenstein says that by using a private language, we can establish a connection between a feeling and a symbol. However, when reliving this feeling, we have to rely on our memory to remember the correct connection. So, there is no way of knowing if we make a mistake. In other words, by not being able to compare our beliefs with those of others, it is impossible to have an idea of what is correct; what seems to be right would indeed be right. Therefore, the statements produced by a lone person wouldn't make sense. And yet this is not merely a communication issue or a problem due to errors in one word or another, because our thoughts are based on the use of language—developed over the course of centuries, through endless social interactions—to consolidate and structure concepts and the relationships among them. The fact that we have language and that we can think and evaluate proposals implies, then, the existence of others—and therein lies the unfeasibility of solipsism.

Dennett and Wittgenstein's arguments demolish the possibility of a Matrix. That said, the conception of reality as a construction of our brain goes hand in hand with the view that we have nowadays in neuroscience. There is no evil genius deceiving us. External reality evidently exists, but in our brain we barely have a very poor and subjective representation of that reality; this perspective is very much in line with Kant's thoughts.

Let's recap the ideas we have discussed in relation to *The Matrix*. With his famous allegory of the cave, Plato doubted the information provided by the senses. Following this line of thought, Descartes later asked if external reality could be the wares of a monumental deception carried out by an evil genius, which eventually led idealists like Berkeley, aware of how easily our senses can deceive us, to deny its very existence. However, these apparent deceptions are no more than the processes our brain must carry out to interpret the information provided by our senses. Maybe the Matrix does have us, after all—though reality is not an illusion created by a powerful artificial intelligence, it is a construction formed by an incredibly sophisticated assembly of neurons: our brain, the real architect of our own Matrix.

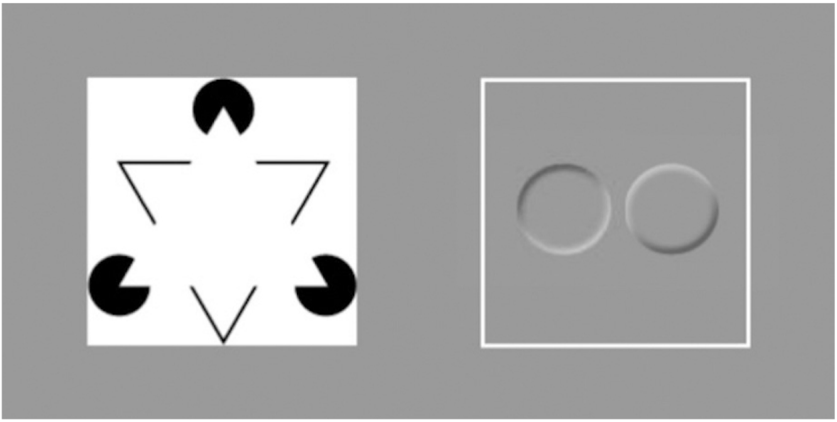
In the late nineteenth century, the prominent German physicist and physician Hermann von Helmholtz (1821–1894)* observed that information captured by our senses is paltry at best and, therefore, the brain carries out *unconscious inferences* to make sense of the external reality. In other words, the brain interprets incomplete information by making predictions based on past experiences.** For example, at any time, we believe we see in detail everything in front of us, but in reality we clearly see nothing more than the space occupied by our thumbnail when our arm is outstretched; we assume the rest.[8]

Helmholtz argued that we do not see copies of reality, of external objects, but rather *signs*, constructions carried out by our brain. These signs don't necessarily have to resemble reality as long as they are reproducible. That means that the representation I have of an object doesn't have to look like the object itself; it suffices that it is the same one when seeing the object on different occasions. As Helmholtz says:

> The objects in the space around us appear to possess the qualities of our sensations. They appear to be red or green, cold or warm, to have an odor or a taste, and so on. Yet these qualities of sensations belong only to our nervous system and do not extend at all into the space around us. Even when we know this, however, the illusion does not cease.
>
> —Hermann von Helmholtz, "The Facts of Perception," 1878

The difference between our perception of reality and reality itself is clearly evident. After all, colors and sounds don't exist in the external space; what exists are electromagnetic waves of different frequencies colliding with photoreceptors in the retina, and variations in air pressure that make small filaments in the cochlea vibrate. Different colors, sounds, smells, textures, and tastes are constructions by the brain, forming our interpretation of the external reality.

Visual illusions give a clear example of how the brain interprets reality through unconscious inferences. The figure on page 91 shows a classic illusion, the Kanizsa triangle, where we infer the shape of a triangle based on its angles and the apparent occlusion of another triangle in the background. Even knowing that the triangle physically does not exist, the perception of the triangle's sides is inevitable. On the right, we see two circles: one appearing debossed and the other in relief, sticking out of the figure's blueprint. In fact, they are the exact same circle rotated 180 degrees. The optical illusion is supplied by the different shadows (above and below, respectively), when assuming that the light comes from above, as dictated by our experience.



Kanizsa triangle and an illusion of debossing and relief.

The inferences described by Helmholtz are what lead us to perceive reality and make sense of what we see. Such inferences are the basis for the functioning of our brain. We use these inferences even in the most mundane circumstances, such as when we play soccer, when we listen to music, or when we watch a movie. In a soccer match, for example, we see where our opponent is heading, where they are looking (and countless other details that we aren't even aware of in the moment), and we predict whether they will make a short or long pass, kick to the center, or aim straight to the goal. When we listen to a familiar song, we predict what's to come based on our memory, but even when we listen to it for the first time, we also make countless predictions because we unconsciously expect the melody to follow certain rules (using notes from specific scales). When we see a movie, a game of sorts takes place between the spectator and the filmmaker: The former tries to guess the plot, and the latter intends for it to be less obvious without veering too far from the spectator's expectations, knowing that a completely unpredictable movie would be unbearable. Alfred Hitchcock, the movie master of suspense, said that an explosion does not cause terror, but its anticipation does. According to Hitchcock, showing a bomb before it explodes creates much more tension than making it suddenly explode.

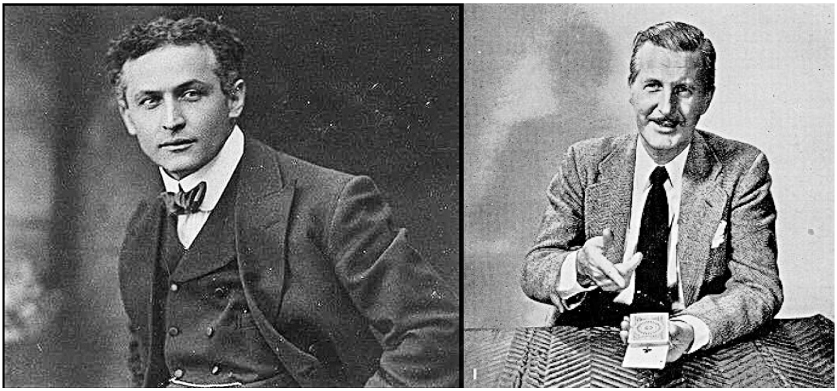
The inferences we are talking about allow us to predict the future and anticipate events, leading us not only to play soccer well or enjoy a Pink Floyd tune or an Ingmar Bergman movie; such inferences are, in fact, vital for our survival. If I see movers hoisting a piano over an upper balcony with just some rope tied to the terrace to brace the load, I imagine what could happen and cross the street; if I see a rabid dog menacingly growling at me, I walk the other way.

At the end of the movie, Neo manages to unravel how the Matrix works; he can clearly see the code that generates the illusion of a hotel hallway and the three "agents" (AI henchmen) he is facing in that instant. As I see it, this is a perfect analogy for how magicians see the behavior of their audiences. These maestros of illusion are the real-world Neos, given their refined perception of how those of us in the audience are continuously making predictions based on past experiences.

Some years ago, I invited Miguel Ángel Gea, a great magician and friend, to give a talk on magic in a packed lecture theater at my university. He began his presentation by saying that the people present were mostly academics, well-educated and intelligent people, and, therefore . . . very easy to fool! Laughter and applause filled the room, but Gea was dead serious. Intelligent people, he continued, are constantly making assumptions about reality that magicians infringe upon with pleasure and gusto in each of their tricks. In fact, performing magic for children is completely different than performing for an audience of adults, because children stop and pay attention to details that grown-ups, over time, have learned to completely ignore.

Magicians tend to avoid repeating the same trick because, by keeping the audience guessing, bombarded with information they can barely process, the spectators don't know beforehand when or where to focus their attention. Harry Houdini, a remarkable escape artist and likely the most renowned figure in the world of magic in the early twentieth century, presented the following challenge: If someone played a trick on him three times, any trick, he would figure out how it was done. In February 1922, a young magician by the name of Dai Vernon—back then completely unknown, although he would later be considered one of the greatest figures in magic—asked Houdini to choose a card, any card, and sign it. He then placed the card in the deck, shuffled it, and—abracadabra—the chosen card was at the top of the stack. He lost the card once again in the deck, shuffled again, and the card once more was on top. Vernon repeated the trick seven times, and Houdini looked on amazed without having a clue how he did it.

What Houdini didn't know was that Vernon was constantly changing the method used to make the card reappear on top of the deck. Each time he thought he was close to figuring out the trick and waited to confirm his suspicions in the next round, Vernon changed the method. In other words, what confused Houdini was making the more than reasonable assumption that if the result was always the same—after shuffling the deck, the card appeared on top—then Vernon was following the same method to carry out the trick.



Harry Houdini (*left*) and Dai Vernon (*right*), two of the most famous magicians in the twentieth century.

For many people, magic is just based on principles such as "the hand is quicker than the eye" or in the fact that the magician makes the spectator look away—for example, by causing a distraction—when he performs the trick. However, the ways magicians perform their tricks are usually more interesting.[9] In fact, many card tricks don't require great speed or impressive sleights of hand and can be done relatively slowly.* On the other hand, distracting the audience by making them look elsewhere can ruin the desired effect: that feeling of not understanding how a card can appear at the top, regardless of having constantly kept your eyes on the deck. Magic's basic ingredient is precisely what Vernon used to trick Houdini: taking advantage of our poor ability to register and process information, and of how we compensate for this limitation by making inevitable and unconscious inferences. It is pure neuroscience, in this case, with a touch of philosophy. The inference Vernon used has its roots in the eighteenth-century thoughts of David Hume, who alleged that if one effect is repeated time and again, it is very hard to avoid inferring that it has the same cause.

The same principle described by Helmholtz about perception also applies to memory. Frederic Bartlett (1886–1969), an English philosopher and psychologist, made his University of Cambridge students read a traditional Native American folktale, "The War of the Ghosts," and then asked them to repeat it to him days, weeks, and sometimes even several years after having read it. Bartlett observed that the retold stories tended to be much shorter and simplified, and that each person modified it according to their personal interpretation.[10] Furthermore, people tended to change the story each time they retold it and, in some cases, after several repetitions, the recap ended up having little to do with the original plot. In other words, more so than the story itself, the students remembered the schema they had made of it according to their interpretation when they read it. Based on this schema, they reconstructed the story in different ways, forgetting many details and unconsciously inventing and adding many others. Bartlett then concluded that memory is a creative process and that its consolidation reinforces a subjective representation, which often changes the memory itself.

There's a big similarity between Helmholtz's construction of *signs* and Bartlett's creation of *schemas*. In one case we refer to vision and in the other to memory, but the process is essentially the same; it implies the construction of a meaning of reality based on unconscious inferences and the use of that meaning, call it sign or schema, instead of reality itself. And just like unconscious inferences give way to visual illusions, similar inferences give way to false memories, that is, to the reinforcement of memories that do not correspond with the original experience.

I briefly mentioned before that for many Argentines the most intense memories are related to soccer matches. In effect, one of my most outstanding childhood memories is that of a match between Argentina and the Soviet Union played during Kiev's winter. I specifically remember this match because of how strange it was to see soccer being played in the snow. I also have a vivid memory of the orange ball. However, funnily enough, that memory is false. I ended up reinforcing in my memory something that never happened: The ball was actually white.

Hugo Gatti, “el león de Kiev” (the Lion of Kiev), catching the (white) ball in Argentina’s match against the Soviet Union.

The way I discovered this memory was a fabrication was quite odd. To the surprise of even their biggest fans, Leicester City—my current town’s soccer team, which up to then was barely able to hold its first division status—became the Premier League champion. Ezequiel Fernández Moores, a fantastic sports journalist, called me in Leicester to ask how we were experiencing what many in England considered to be the most surprising result in the history of soccer. As we spoke, one thing led to another, and the memory of the orange ball in the match against the Soviet Union came to light. However, one of Ezequiel’s colleagues, who had an encyclopedic memory, corrected me and said the ball was actually white and, in fact, there was a famous photo of that match where Hugo Gatti, “el León de Kiev,” is shown grabbing the ball in midair (see photo). My confusion happened because some time later, Gatti was asked to choose the ball that would be used in the famous Superclásico match between Buenos Aires rivals Boca Juniors and River Plate. Gatti assumed the field would be covered in the confetti that fans often threw when their team came out (as in the 1978 World Cup), so he asked for an orange ball. The memory of that orange ball was then consolidated in my memory by making an incorrect unconscious inference.*

Why do we register and remember so little? With billions of neurons in our brain, couldn’t we process and store a much more reliable representation of reality? Think back to the discussion in the previous chapters where we compared the functioning of our brain with that of computers. The key is that the brain does not seek to create a detailed reproduction of reality and instead focuses its resources on trying to understand it. Contrary to a computer, which can keep thousands of photos, songs, and high-resolution videos in its hard drive, what we see or remember is actually quite little. However, we do understand what happens to us and the computer doesn’t.

In the short story “Funes the Memorious,” Jorge Luis Borges describes the misfortunes of a gaucho from Fray Bentos who, after receiving a blow to the head, acquires the incredible ability to remember absolutely everything: “Funes not only remembered every leaf of every tree of every forest, but also each time he had perceived or imagined it.”[11] Funes ends his days bedridden in his dark room, awash with memories and irrelevant details that didn't even let him think. With Funes's story, Borges shows with incredible clarity the importance of oblivion and how tortuous it would be to remember everything. In fact, the author concludes that Funes wasn't very capable of thinking because "to think is to forget differences, to generalize, to abstract."**

As has been said, in an area of the brain dedicated to memory, we found neurons representing concepts (Jennifer Aniston, Luke Skywalker, and so on). These neurons do not code any of these people's details—how they're dressed, if they're face forward or looking to the side, and so on—precisely because we tend to forget details and remember concepts and associations between them. From these concepts we create Bartlett's schemas, the subjective representations of how we remember events of our lives.

We have said that the idea of *The Matrix* is impossible, although it offers a very useful analogy for how the brain works and how we create a representation of reality. Yet, the movie's plot and the idea of being immersed in a great simulation is actually spine-tingling, considering how we live connected to the internet and how our day-to-day reality is constantly bombarded with information, be it via email, WhatsApp, Messenger, hundreds of TV channels, cell phones, or tablets. Such a barrage does little more than turn us into Funes, devouring extremely valuable time that would be best used wondering about countless other things. Likewise, our social lives increasingly depend on mechanisms for virtual interactions, such as Facebook or Twitter, and as if we are one of the many people experiencing the Matrix's deception, we're addicted to computer games and we spend hours in front of the TV to vicariously experience the lives of strangers: In minutes, we go from being action heroes to martial arts experts, secret agents, or bank robbers. Ultimately, our lives unfold in front of one screen or another, and we are less and less ourselves. So, let's take the red pill, turn off the computers, disconnect our cell phones, and start living our real lives.

---

* Neo has stored the disk inside a book titled *Simulacra and Simulation*. The recurrent theme in that book, published in 1981 by French philosopher Jean Baudrillard, is the vague difference between the real and the imaginary, or that which is a simulation. For example, Baudrillard contrasts what is presented as imaginary in Disneyland and the (apparent) reality of a city like Los Angeles (anyone who has lived in LA perfectly understands what he is referring to). The confusing difference between reality and illusion is *The Matrix*'s main theme. The Wachowski siblings, the movie's screenwriters and directors, actually made the actors read Baudrillard's book before filming. The movie also makes a clear reference to Lewis Carroll's *Alice's Adventures in Wonderland* by having Neo follow a white rabbit.

* For Berkeley, just like *esse est percipi* (to be is to be perceived) was applied to things, *esse est percipere* (to be is to perceive) was applied to the mind.

** Bishop Berkeley also solved many of the contradictions in his ontological idealism by appealing to the existence of God. According to Russell, if Berkeley had been consistent in his doctrine, he would've had to deny any possible knowledge of God and would have ended in solipsism (Bertrand Russell, *History of Western Philosophy*, p. 564).

* Hilary Putnam, *Reason, Truth and History* (Cambridge: Cambridge University Press, 1981). Based on logical arguments that have been thoroughly discussed in contemporary philosophy, Putnam claimed it is impossible for us to be a brain in a vat, refuting in this way the Cartesian skepticism in which reality is an illusion created by an evil genius. For more details on Putnam's arguments, see chapter one of his book. In the following pages, I will present a much simpler and more direct rebuttal proposed by Daniel Dennett as well as the one put forward by Wittgenstein based on language.

* However, the evocation of memories in these patients worked only in 8 percent of the cases, and in almost half of them, the memories were the same as the hallucinations they suffered during their epileptic seizures. Therefore, it can also be argued that more than triggering the flow of consciousness, Penfield simply evoked epileptic seizures. However, beyond these possible interpretations, the fact that he could evoke memories through electric stimulation in the brain is nothing less than spectacular.

* Wittgenstein's arguments are not trivial at all and have led to multiple interpretations by different philosophers. My brief description probably does not do justice to the magnitude of his ideas. For a more detailed description, see, for example, chapter twenty-two of Anthony Kenny's *Brief History of Western Philosophy* (see endnotes). Or, even better, as Dennett says, one can certainly ignore Wittgenstein's interpretations by other philosophers and read his original texts instead (Dennett, *Kinds of Minds*, p. 170).

* The variety of contributions by Helmholtz is truly striking. Among other things, he formulated the principle of energy conservation and posited the notion of free energy in thermodynamics; he invented the ophthalmoscope, measured nerve conduction velocity, introduced a mathematical description of acoustic vibrations, and laid the foundation of the current color theory, using three variables to define it (hue, saturation, and brightness).

** Similar arguments were also previously made by Ptolemy, Aristotle, Alhazen (a.k.a. al-Haytham) and Descartes, among others. For more information, see chapter nine in *Borges and Memory* (cited in chapter one).

* Case in point, one of the most remarkable Argentine magicians, René Lavand, used time and again the catchphrase "It can't be done slower" when manipulating the cards with his only hand. When I visited the Magic Castle in Los Angeles, the mecca of magic, I was pleasantly surprised to see an enormous portrait of Lavand in one of the main rooms. That's where I discovered that this man, whom I knew from TV shows in Argentina, was considered one of magic's luminaries worldwide.

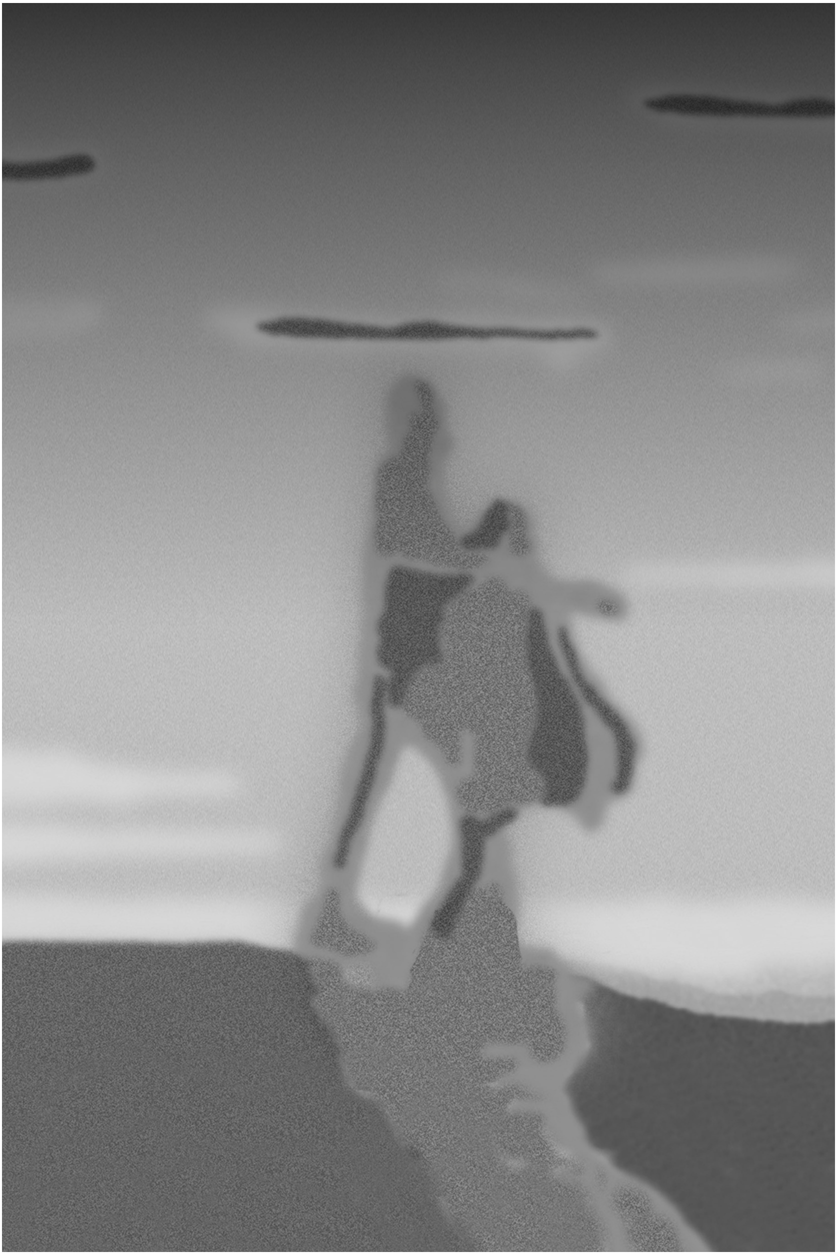
* In fact, I don't think I'm the only one with the false memory of the orange ball in Argentina's match in the snow. When searching for proof of this, I saw an interview with Beto Alonso (who killed it in the Boca match) where he said something similar, and another in the newspaper *Clarín* (from February, 7, 2013) that also claimed the ball in the match against the Soviet Union in Kiev was orange. However, the photo of "el León de Kiev" catching the white ball is irrefutable. For a more detailed description of false memories, see chapter four of *The Forgetting Machine* (cited in chapter two of this book), or the works of Elizabeth Loftus, an expert on the subject, among them Elizabeth F. Loftus, "Creating False Memories," *Scientific American* 277, no. 3 (1997): 70–75.

** The importance of oblivion in our thoughts is the prevailing theme in *Borges and Memory* (cited in chapter one).

CHAPTER 5

UNTIL THE END OF THE WORLD

Reading the Mind



Together with Werner Herzog and R. W. Fassbinder, Wim Wenders, director of *Until the End of the World*, is one of the great exponents of the New German Cinema of the 1970s. Reflecting his passion for photography and rock, Wenders's 1991 movie combines music by U2, Lou Reed, and Peter Gabriel with magnificent landscapes in the Australian desert, and deranged shapes and colors from downtown Tokyo. It is a road movie of sorts, filmed in ten different countries, rounded off by a sci-fi finale in an isolated lab in the middle of the desert.

The movie's scattered plot was unanimously destroyed by critics, perhaps due to the expectation created by Wenders's previous movie, *Der Himmel über Berlin* (known in English as *Wings of Desire*), shot in 1989 in a spectacular Berlin, shortly before the fall of the wall. *Wings of Desire* is pure poetry, melancholy, philosophy, and humanism. It opens with the poem "Lied vom Kindsein" by Peter Handke, Wenders's co-screenwriter, who in the fourth verse sets the movie's tone and depth:

Als das Kind Kind war,
war es die Zeit der folgenden Fragen:
Warum bin ich ich und warum nicht du?
Warum bin ich hier und warum nicht dort?
Wann begann die Zeit und wo endet der Raum?
Ist das Leben unter der Sonne nicht bloβ ein Traum? . . .
Wie kann es sein, daβ ich, der ich bin,
bevor ich wurde, nicht war,
und daβ einmal ich, der ich bin,
nicht mehr der ich bin, sein werde?

When the child was a child,
it was the time for these questions:
Why am I me, and why not you?
Why am I here, and why not there?
When did time begin, and where does space end?
Is life under the sun not just a dream? . . .
How can it be that I, who I am,
didn't exist before I came to be,
and that, someday I, who I am,
will no longer be who I am.

Two angels, Damiel and Casiel, glide over Berlin's streets, drawing close to different people, listening to their thoughts and trying to understand their lives and what they feel. Lacking the ability to feel for themselves, they get together daily to exchange experiences they have witnessed: the lives of children, parents, grandparents; men on their deathbeds;

inconsolable adults; and many other people trapped in their daily routines. The movie is a nearly unbroken stream of consciousness—the consciousness of those whose inner voices the angels tune in to like a radio dial—comparing the melancholic everyday life of humans with the angels' eternal longing to feel for themselves.

In a genius aesthetic play, Wenders juxtaposes scenes in color, representing the humans' point of view, with others in black-and-white, showing the angels' view of reality, incapable of sensing colors. Damiel confesses to Casiel that he longs to feel what it means to come home, take off his shoes under the table and stretch his toes with bare feet; eat roasted lamb with wine; have a fever; have ink-stained fingers after reading the newspaper; to only have suspicions rather than knowing it all. Such is his weariness that when the opportunity presents, he decides to resign his immortality to become human.

Victory Column (*Siegessäule*) in Berlin, where two angels, Damiel and Casiel, gather to share their experiences among humans in *Wings of Desire*.

When he wakes as a man, he walks, elated, and feels the weight of his body in each step; he crosses paths with a passerby and asks him about the different colors on the street because, although since the dawn of time he has perfectly known what colors are and all they represent, he can finally see them. He orders a coffee, grabs the cup in his hands, feels the heat, and, for the first time, he stops to smell coffee's particular aroma before savoring it.

Damiel's story has a lot in common with an archangel imagined by British philosopher Charlie Dunbar Broad.[1] According to Broad, this archangel possesses endless chemistry knowledge and knows, for example, each and every detail of the microscopic composition of ammonia and how it behaves in countless chemical reactions. But just like Damiel did not know the smell of coffee before becoming human, Broad argues that this archangel, beyond all his knowledge of chemistry, couldn't even begin to imagine the penetrating smell of ammonia. In other words, the knowledge of material processes cannot explain how we sense reality.

Let's continue a bit further with *Wings of Desire* because it will help us refresh the discussion about materialism and, from there, the possibility of reading the mind. The nostalgic black-and-white shots of Berlin dominate Wenders's visual narrative; colors mainly appear toward the end of the movie, when Damiel becomes human. This actually brings to mind another great discussion in contemporary philosophy: the story of Mary, a scientist imagined by Australian philosopher Frank Jackson.[2]

Mary is an expert in visual perception who spends her life locked up in a lab where absolutely everything is black-and-white: the set of instruments, the books, the computer monitor, the TV. Like Damiel, Mary never saw colors, although she perfectly understands what they mean and the activations they generate in the brain; Mary knows everything there is to know about the perception of colors. But one day she leaves the lab and, like Damiel when he became human, sees colors for the first time. Will she be surprised by the colors she sees? If she already knows everything there is to know about them, what new information will she glean from being able to sense them?

This is what is known as the *knowledge argument* to reject materialism. When Mary leaves the room, she could, in principle, be learning something new by seeing the red of an apple for the first time. But if she already knew absolutely everything about the physical processes that unfold when seeing different colors—that is, the neuron responses that they generate—and yet feels something new when initially seeing colors, then the mental processes must involve information that goes beyond the material. Jackson concludes that the mind cannot be reduced to physical processes and that materialism is false.

Similar arguments were previously made by American philosopher Thomas Nagel, when he tried to imagine how would it feel to be a bat.[3] This is quite far from imagining someone like Dracula. Nagel's posit is very different: He doesn't ask himself what it would be like to be a man living like a bat, but rather what it is like to be a bat. How do their perceptions of reality and their thoughts work?

The bat was not an arbitrary choice, since they have limited sight and perceive their surroundings, for example, in a completely dark cavern, using the echo produced by the sounds they emit. Nagel says that even if we could understand every detail about how a bat's echolocation system works, we will never be able to imagine what it is like to be one of them. However, contrary to Jackson, Nagel doesn't use this argument to explicitly reject materialism, but rather to illustrate the difficulty we have in understanding the relationship between the physical processes (that is, our knowledge of how echolocation works) and the mental ones (what we would feel if we were perceiving the echoes). This is precisely the hard problem we introduced in chapter two.

Let's get back to Mary and Jackson's apparent rejection of materialism. To Dan Dennett, the brilliant philosopher previously introduced, Jackson's argument is based on an intuition error. If we follow Jackson's premises exactly—argues Dennett—then there would be no problem at all.[4] The error lies in the fact that it is difficult to intuit what it really means to say that Mary knows *absolutely everything* about the perception of colors. If she, in effect, knew it all, she would know how neurons in the visual cortex activate and how they would, in turn, transmit this information to hundreds of other areas, triggering memories, emotions, and countless conscious and unconscious processes. In other words, Mary would know beforehand how she would react when perceiving the red of an apple and, when seeing it for the first time, she wouldn't be surprised at all, given that her reaction would be exactly what she expected, an accumulation of processes that lead to qualia, the feelings produced when seeing the red of an apple, the brown of coffee grinds, or the white of jasmine flowers.

Dennett argues that we have a similar problem when trying to imagine what it is like to be a bat.[5] Once again, it is an intuition error and not a philosophical dilemma. In this case, the problem stems from the difficulty we have in imagining how to perceive our surroundings with an echolocation system, interpreting what unfolds around us as if using sonar. Let's assume, for example, that such a system has only a few yards of resolution in bats. So, what would we feel then, as bats, when attacked by an eagle? Until the last few seconds, we would feel absolutely nothing, because we wouldn't even notice the danger of the eagle approaching. And what would we feel in those last few seconds? This will likely depend on our ability to establish, with our sonar system, if the eagle is heading our way, if it has its claws ready to attack, and so on. Based on these arguments, Dennett concludes that regardless of how anti-intuitive it may be, when acquiring information on how a bat's brain works, we can effectively begin to understand what it is like to be one of them.

In conclusion, materialism firmly resists the philosophical attacks of bats and scientists unable to see color. After identifying and setting aside deceitful intuitions, it is clear there is nothing magical or mysterious that takes place in our mind; it is simply the activity of billions of neurons. Having clarified this point, we can now face the big question of this chapter: If the mind is nothing more than the functioning of the brain, can we read the mind? Can we reconstruct what a person is seeing, feeling, or thinking from the activity of their neurons? Clearly, we wouldn't hear, like Damiel and Casiel, the minds of those around us just by paying attention to them. Instead, we would need to use sophisticated measurements of the brain's activity and complex algorithms to interpret what such activity represents. This is exactly the plot in *Until the End of the World*.

Art is extremely subjective. As I mentioned at the start of this chapter, critics concertedly trashed *Until the End of the World*. Although, in Wenders's defense, his final version lasted four and a half hours, but since it was impossible to commercialize, it was reduced to two and a half hours, as required by the producer's contract. Despite the absence of some narrative coherence, *Until the End of the World* continues to be one of my favorite movies. Perhaps that is due to its soundtrack, or the spectacular scenes it depicts, but I believe, above all, it's due to the subtleties of the final part of the plot.

The film could be divided into two somewhat disconnected parts. The first is a road movie where one man travels frantically around the world, carrying a special camera that records not only the images he films but also the cerebral reactions of the people being filmed. In the second part, he arrives at a faraway place in the Australian desert. There he meets his father, a scientist who has built an impressive lab with the exclusive aim of helping his blind wife to see. To that end, he plans to directly transmit to her brain the neural activity recorded during the filming of the different scenes. The process is exhausting, but after countless adjustments, the woman slowly begins to recognize blurry scenes (like the one illustrated at the opening of this chapter), as her brain interprets the signals she receives. Yet once this is achieved, in and of itself a great feat, the scientist plans to carry out the reverse process: instead of transmitting to the brain the images recorded by the camera, he tries to interpret the activation patterns naturally produced by the brain. And that's how he begins to read the mind after *decoding* the information provided by the cerebral recordings, projecting thoughts on a screen and even catching glimpses of thoughts during sleep. The plot obviously lacks a scientific foundation; after all, it is merely fiction, although, surprisingly, not so much anymore.

Almost twenty years after first seeing the movie, fate had me doing something similar: Based on the responses to images of Jennifer Aniston, Halle Berry, and other people (described in previous chapters), I asked myself if, like the scientist in *Until the End of the World*, I could manage to reconstruct what the patient saw. (Remember, in epileptic patients, once intracranial electrodes were implanted to find the epileptic focus, we recorded neurons that responded to pictures of different people). The idea is simple: If the Jennifer Aniston neuron fires, I can tell that the patient is seeing a picture of Jennifer Aniston. If the Halle Berry one fires, I can guess, instead, that the person is seeing Halle Berry, and so on. When considering more neurons and things to predict, the job becomes more complex, but we could solve this using decoding algorithms. Setting aside technical details, the decoding algorithm makes an informed guess of what image leads to the observed responses in a group of neurons.[6] The result was that, in effect, we were able to guess, much better than by chance, what picture the patient was seeing at each moment.[7]

Sometime later we got interested in an even more fascinating question: If we could predict what the patient is seeing, could we also project what they are thinking on a screen, like in *Until the End of the World*? Of course, we need controlled conditions. We can't predict and then project any random thought, but we could do so with thoughts involving people for which we have neurons responding. For example, if we see that the Jennifer Aniston neuron fires, we project her image on the screen; if a Marilyn Monroe neuron fires, we instead project her image, and so on. But how do we make one or another neuron fire? Easy, we ask the patient to think, one at a time, about the people that these neurons respond to. Then we use decoding algorithms to predict what the patient is thinking, and finally project this image on the computer monitor. Just like in the movie, where the scientist can hardly contain his excitement when seeing how, slowly but surely, images from the brain were projected onscreen, sometime later, in a real lab, we were astonished as we saw that the patients could voluntarily control the firing of their neurons and project their thoughts on a computer screen.[8]

Let me be clear, we did nothing more than project what the patients voluntarily wanted to show us. This required a certain amount of effort and focus on their part, thinking about one or another person, so that our mathematical algorithms could decode and later project the image of these people on the screen. We are very far from peeking around private thoughts. On the contrary, we could say that we developed an alternative communication method: instead of telling us what they were thinking, the patients voluntarily projected these thoughts. Why is this important? First, because it shows the ability a person has to voluntarily modify the firing of their neurons through thought. Once again, we reject Cartesian dualism; we know the mind and the activity of neurons are the same thing. Second, because this technology could eventually be used on patients with *locked-in syndrome* (those who are condemned to being locked inside themselves, without the ability to communicate or move at all), or other patients with severe motor deficits (like the famous physicist Stephen Hawking), who have limited motor abilities and can hardly communicate, be it by speaking, writing, or even moving a finger. So, it is possible that in a few years these patients might be able to communicate directly with the external world through their neural activity.

A particularly dramatic case is the one of vegetative patients. The vegetative state is often confused with a coma, but it is not the same. While a patient in a coma is sleeping and unconscious, someone in a vegetative state can be awake, gesticulate, open their eyes and stare as if they are looking around even though, in truth, they can't see or feel anything, since they are not conscious. The main characteristic of the vegetative state is that the patient doesn't voluntarily respond to any commands (for example, asking them to squeeze a hand, blink, voluntarily move their eyes, and so forth). In some cases these patients recover in a few days or weeks, but sometimes they remain in that state for years, artificially fed (usually they can breathe on their own) and with minimal chances of recovery. It is especially in these cases, after years of isolation, that we wonder if they are truly not aware of anything, or if perhaps they are but, like those with locked-in syndrome, they can't make the smallest move to show it.

Let's imagine, for an instant, the despair of a bedridden person who is conscious of what is happening around them but can't let anyone know. Do they want to tell us something specific? Do they want to listen to music, turn off the light, or ask us not to be so noisy? Do they feel pain and perhaps want to scream aloud so that we raise the doses of painkillers? (This reminds me of the 2007 movie *Awake*, where a patient discovers after being anesthetized that he is still conscious and suffers pain throughout the entire surgical procedure without being able to even move a muscle for anyone to take notice.) Such considerations also lead us to raise a more dramatic question: After several years in those conditions, would a person want to continue living, considering the almost negligible chances of recovery?

The most well-known of these cases is that of Terri Schiavo, who suffered a heart attack in 1990. Paramedics were able to revive her, but given the long absence of blood flow to her brain, she remained unconscious and, several months later, was ruled to be in a permanent vegetative state.

In 1998, after years in the hospital surviving through a feeding tube and without any response that showed even a minimal degree of consciousness, her husband asked for her to be disconnected from life support devices and allowed to die, since he argued that that would have been her wish if she were aware of the situation. The request was rejected by Schiavo's parents, which began a long legal dispute involving different members of Congress, the Supreme Court, the intervention of then Florida governor Jeb Bush, and even of his brother, President George W. Bush. Only in March 2005, after seven years of several judicial rulings and appeals, Terry Schiavo died when the feeding tube that kept her alive was disconnected.

The legal journey triggered by Schiavo's case sparked heated public debate. What should we do in these cases? Is using artificial methods justified to prolong someone's life for years, even though they don't have a chance of regaining awareness? To a large extent, the legal dispute over Terri Schiavo's case was based on interpreting what she would have wanted in those circumstances: While her husband said that she would have liked to go in peace, her parents claimed that, as a fervent Catholic, she would have chosen to continue her life, even bedridden as she was. Assuming she could still have been conscious, we then ask ourselves if there was any way we could have read her mind and truly interpreted her will. In fact, the judicial decision on her fate was in part delayed because the judges believed that new scientific advances would allow communication with her.

In September 2006, a year after Schiavo's death, an article in *Science* magazine produced even more stir.[9] The study was led by Adrian Owen, a British neuroscientist who has worked on cases like Terri Schiavo's. His job is Herculean and, at first, frustrating: He develops sophisticated methods to determine if patients are still conscious and, if so, to break the barrier of their complete immobility and prostration to eventually manage to communicate with them. To do so, he uses functional magnetic resonance imaging (a.k.a. fMRI, a method that produces images of the brain and shows which areas activate when carrying out certain tasks), exploiting the principle that both the execution and the imagining of certain tasks produce similar stereotypic activations in the brain. For example, the execution or imagination of a movement activates areas involved with the processing of motor tasks.

The patient that Owen and his colleagues analyzed was a woman who had remained in a vegetative state after a car accident. The woman spent several months without voluntarily responding to any command or showing any sign of being conscious. However, when Owen asked her to imagine she was playing a tennis match while doing an fMRI recording, he noticed a clear activation of the motor area in the cerebral cortex, and when he asked her to imagine she was walking around her house, going from one room to another, he saw activity in the areas related to spatial navigation, the same areas that are normally active in other people.

The follow-up step is clear, but no less amazing. If the act of imagining different tasks produces specific activity in the brain, this could be used as a communication method with patients in a vegetative state: We simply ask them to imagine they are playing tennis when they want to say yes, or imagine walking around the house when the answer is no. This is exactly what Owen and his colleagues did and showed in an article published a few years later.[10] With a patient (referred to as Patient 23) who had been in a vegetative state for five years, they managed to have the following dialogue:

Scientist: Is your father's name Thomas?
Patient 23: No.
Scientist: Is your father's name Alexander?
Patient 23: Yes.
Scientist: Do you have siblings?
Patient 23: Yes.
Scientist: Do you have sisters?
Patient 23: No.

After years of being unable to communicate with anyone, Patient 23 correctly answered five consecutive questions. Why didn't they continue asking more? Because in order to obtain reliable answers, the patient has to focus on imagining the action of the response for about half a minute and, moreover, the researchers had to repeat the question several times to validate the answer. Obviously, both Owen and the patient's relatives would've liked to have a more fluent dialogue and ask countless questions. Although it is true that the interaction up until now is limited, these results are a clear demonstration that it is possible to read the mind of someone in a vegetative state.

The ability to ask more questions or have more elaborate interactions is mainly a matter of technology, optimizing signal-processing algorithms, the types of questions asked, and so forth. The bad news is that Owen's team only managed to communicate with about 20 percent of their subjects. Perhaps with technological advances, in the future it might be possible to increase this number, but just the fact of being able to read the brain of one out of five patients in a vegetative state is in itself a major feat. Although it is also important to note, to avoid generating false hope, that this procedure is not applicable to all patients in a vegetative state, since in many cases the lesions, involving primary and vital structures in the brain, are irreversible and there is no possibility at all of communicating with that person or expecting that they could eventually regain consciousness.

The big dilemma is still looming: Would these people wish to continue living in those conditions? Owen argues (in my opinion, correctly) that there is still a long way to go before raising these types of questions. It is not just a matter of technology or certainty in the interpretation of the answers, but also of thinking about what to do if the individuals eventually answer no.

Clinical relevance notwithstanding, let us temporarily set aside the extreme case of patients in a vegetative state, where the possibility of reading their mind is crucial, although relatively limited, to ask ourselves if we could extract much more sophisticated information by applying similar methods to healthy subjects. We already saw that through the implantation of electrodes in epileptic patients, we can register the activity of dozens of neurons and predict what they are thinking, and even project their thoughts. The question then becomes whether it is possible to do something similar with methods that don't involve a surgery or the implantation of electrodes in the brain. Of course, the brain signal measured from outside of the head (without any surgical intervention) is much weaker and less precise than the one obtained when directly recording neural activity with intracranial electrodes. Although perhaps the development of more elaborate methods for processing these signals will allow us to go beyond obtaining only affirmative or negative responses, as with the vegetative patients. Ultimately, we're not satisfied by obtaining answers to predetermined questions and, like the scientist in *Until the End of the World*, we must ask ourselves if it is possible to directly access a person's thoughts.

A group of Israeli researchers led by Uri Hasson found something amazing in this respect.[11] While they showed a group of subjects the 1966 spaghetti western *The Good, the Bad, and the Ugly*, scanning the subjects' cerebral activity with fMRI (the same type of recordings used by Owen with patients in a vegetative state), they found that certain scenes produced the same specific activations in all subjects: The shots with faces activated an area called *fusiform gyrus*, those with landscapes activated a nearby area, and shots showing the manipulation of objects with hands activated an area farther away.

The activation of these three areas in the brain then allowed the prediction of whether the person was seeing faces, hand movements, or landscapes. The big difference from the results obtained with patients in a vegetative state is that, in this case, the researchers didn't have to ask the subjects to focus for half a minute (and several times) on a specific action. Instead, the activations, marking the presence of one shot or another, were produced spontaneously while the subjects watched the movie.

When considering the possibility of reading someone's mind, the most important limitation in this work is due to the fact that the number of areas that are selectively activated when seeing different types of images, or doing certain tasks, is quite limited. Owen and his team wisely chose tasks—playing tennis or walking around the house—that activate different areas of the brain. Similarly, Hasson's group distinguished faces from landscapes and scenes with hand movements. However, we must ask how many more things could be identified through the activation of specific areas of the brain?

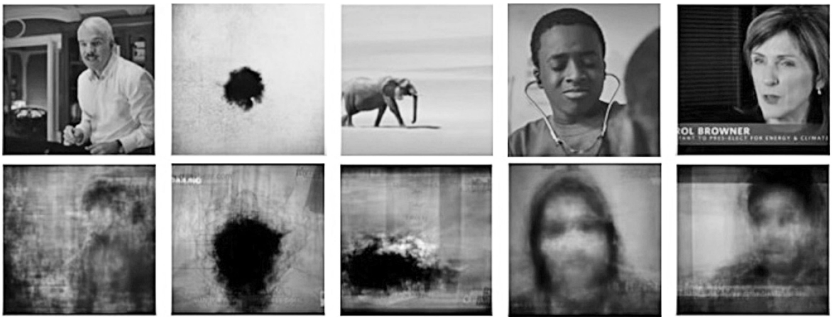
Just like the images we see on a computer monitor are made up of pixels, those obtained with fMRI are made up of *voxels*. To simplify this, we can roughly equate each voxel with a pixel, whose intensity represents the activation of a tiny part of the brain measuring a couple of millimeters. Groups of adjacent voxels compose the areas previously mentioned, with a span measured in centimeters. The work of Owen's and Hasson's teams, like many others in neuroscience, is specifically based on quantifying the average activation of the group of voxels that are part of different areas.

Analyzing the average activations in distinctive areas of the brain has been the standard method used to interpret fMRI recordings. However, a significant revolution in the analysis of this data unfolded in the early part of this century when American scientist Jim Haxby, together with a group of colleagues, asked himself if the representation of faces and diverse objects is given by the activation of specific areas in the *visual cortex*.[12] The experiment consisted of showing pictures of human faces, cats, and different objects—houses, chairs, scissors, shoes, and bottles—while recording the activity of voxels in the visual cortex. When analyzing the results, Haxby and his colleagues didn't find a specific area responding to shoes or another that was activated by different types of bottles, but they did see that different stimuli activated the same visual areas, although not in the same way. This is what is technically known as a *distributed and overlapping* representation of the stimuli. Despite not activating specific areas, based on the activation pattern of the voxels in visual areas, they were able to predict with a very high degree of certainty the stimuli that the subjects were seeing.

The great revolution came through the methodology they developed to reach this result. The details of the method aren't as relevant; what is important to understand is that they analyzed the activation pattern in each of the voxels within the visual areas, rather than the average activation of one area or another, which is what was being done until then. Notably, these results do not deny that there are areas that are activated with faces or landscapes, as discovered by Hasson's group. On the contrary, Haxby's results show that beyond these large-scale responses involving the combined activation of voxels in specific areas, there are also activations on a much finer scale, which could be observed when considering each voxel separately rather than just looking at them as part of a specific area.[13]

Several years ago, I was invited to give a talk at a conference in Maine. I can't quite recall what I spoke about, but I do vividly remember the impact that a presentation by Jack Gallant—a neuroscientist from the University of California, Berkeley—had on me. Gallant's group was already obtaining astonishing results refining the method developed by Haxby to extract information through voxel activation patterns, for example, predicting which of 120 images a subject was seeing.[14] Yet the results he presented in Maine were even more amazing. In this case, Gallant's team first recorded the brain's activity—of thousands of voxels—while the subjects watched hours of videos: scenes from movies, documentaries, landscapes, and so forth; then the researchers created a computer model describing the characteristics of the scenes that activated each voxel; and finally they showed the subjects new videos (that is, videos that weren't used to create the model), and based on the activation of the voxels, they then used the model to reproduce what the subjects were seeing at each point in time.[15]

Like the rest of the scientific community, I was quite amazed by the results, but in my case the surprise was truly profound. After all, Gallant was carrying out, in real life, the experiment of the scientist in *Until the End of the World*. The results are far from perfect, but the method reproduced, at least in general, the shapes of the things that the subjects were seeing and, curiously, these reproductions are similar to those blurry images used by Wenders in his movie.



Scenes from the videos shown to the subjects (*top*) and reproductions of these according to their brain activity (*bottom*). (Adapted from Shinji Nishimoto et al., *Current Biology*, 2011; see chapter endnotes.)

So far, we have shown that it is possible to predict what someone is seeing through the activation of voxels measured by fMRIs. These results are amazing, but they are quite far from what we commonly understand as "reading the mind." So, will we really be able to read thoughts using this technology? That is, will we be able to not only reconstruct what a person is seeing but also what they are thinking, or even dreaming?

We described how Adrian Owen managed to communicate with patients in a vegetative state by asking them to imagine different tasks, which produced similar neural activations to those obtained when carrying them out. Other experiments also showed that the same areas of the brain are activated by both seeing and thinking about faces or landscapes.[16] Based on these results, it then seemed possible to predict what a person is spontaneously thinking, without needing to show them any type of stimulus.

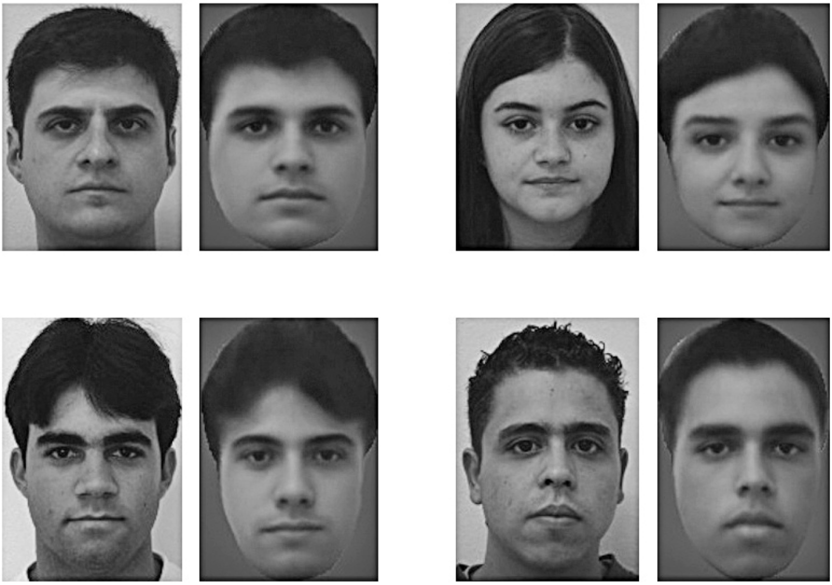
One of the most interesting aspects of *Until the End of the World* is Wenders's description of the fascination people felt when seeing images of their dreams. But will we manage to read someone's dreams in real life? Although it seems impossible, this is exactly what Yukiyasu Kamitani did with his team of researchers at Kyoto University.[17] First, they showed subjects a large number of photos while recording their brain activity with fMRIs, so as to determine what each voxel was coding, and then to predict what the subjects were seeing. Nothing new yet. Kamitani and his team's brilliance came when they later put the subjects to sleep and continued recording their brain activity to see if they could predict their dreams. But how could they verify they were really interpreting the subjects' dreams and not just anything else? Simple: They woke the subjects up regularly, asked them what they were dreaming about, and later compared their answers to the predictions made through the brain activity measured shortly before waking them. The researchers got the dream content right in 60 percent of the cases, which is far from perfect, but significantly better than what they would have obtained by making random predictions.

We've shown that it is possible to predict what a person is seeing, thinking, or even dreaming using brain activity measured with fMRIs. And additionally, that it is possible to reproduce images that a subject is seeing through voxel activity. These reproductions are certainly limited, only discerning the shapes and contours of what is being shown rather than specific features of a face. However, this is still surprising given the low resolution of fMRI: Although it is true that we obtain much more information when considering the voxels individually and not as part of one or another area, one voxel represents the combined activity of millions of neurons and it doesn't deliver information about what each neuron is doing. It is like being outside a stadium and trying to interpret what is happening by listening to people bellowing at a distance, without discerning what each person is saying. How much further would we get if we could distinguish the activity of the neurons that make up each voxel? Unfortunately, this is not possible with fMRI, but it is possible when implanting electrodes in the brain.

Many of science's great ideas are relatively simple. Usually, what is surprising about these ideas is not so much their intellectual complexity, but rather that no one else had thought about them before. A few years ago, Doris Tsao and Winrich Freiwald had such an idea. First, they made fMRI recordings in monkeys, while showing them hundreds of faces and other objects; then they identified six specific areas that responded to faces (which nowadays are known as *face patches*); and last, they introduced electrodes to register the neural activity in those areas, showing that almost all neurons in the face patches responded mostly to faces.[18] In a series of spectacular experiments, Tsao and Freiwald have been interpreting the specific functions of these face patches in coding faces.[19] The big difference with previous works is that in those cases the researchers introduced electrodes in large swaths of the visual cortex, hoping to find one or another response to faces, whereas Tsao and Freiwald knew exactly where to implant the electrodes, noticeably increasing the probability of finding neural responses.

The most remarkable result came a few months after I first wrote these lines. After several years of research, Tsao found the exact model that was close to perfect at characterizing these neural responses. In other words, she identified what facial features each neuron responded to, be it the width of the lips, the length of the nose, the separation between the eyes, or more complex characteristics. Two truly amazing predictions emerged from these results. First, once the model for what these neurons responded to was established, it should, in principle, be possible to predict their response to any face, just by introducing the face's characteristics in the model, which would calculate the neurons' responses. Second, based on a group of neurons' responses, it should be possible to carry out the reverse process and reproduce the observed face, with much more precision than the one obtained through fMRI. This is exactly what they accomplished.[20]

The similarity among the faces shown and those reproduced is extraordinary. At the risk of sounding redundant, I highlight this result because it is truly incredible: The reproductions in the images shown on page 117 were obtained by only considering the activity of two hundred neurons, using a model that combined, in the right proportion, the different characteristics—nose length, mouth size, and so forth—coded by each one of them. Basically, it is like they were using the model to build an identikit of the faces based on the firing of the neurons.

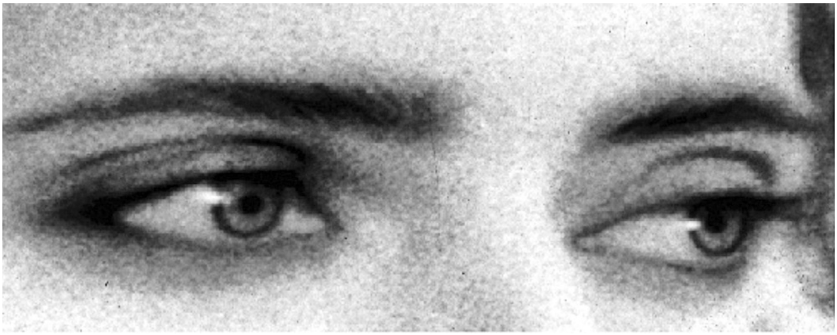


Faces shown to a monkey (*left*) and reproductions of these faces made based on the activity of a few hundred neurons in face patches (*right*). (Adapted from Chang and Tsao, *Cell*, 2017; see chapter endnotes.)

More than one person might think that these results with monkeys, like those with fMRI or the ones with epileptic patients, where we had them project their thoughts on a screen, are fascinating yet also somewhat controversial. It is scary to think that we can reconstruct what a subject is seeing or thinking by measuring their brain activity. However, we cannot carry out predictions about what one subject thinks without them agreeing to collaborate with these types of experiments, in order to adjust the algorithms to be used, and so on. We are very far from wandering the streets, like Damiel and Casiel, and interpreting the private thoughts of passersby.

Nonetheless, beyond this exception, we are somehow constantly reading other people's thoughts; that is, we may not know if a person we are talking to is thinking about their future Caribbean vacation or Sunday's soccer match, but we can interpret some of their emotions. We do it all the time and so naturally that we are typically not aware of it. We don't analyze neural or voxel activity, but we do scrutinize the tiny, almost imperceptible, unconscious gestures of those with whom we interact to decipher their mental state; if they seem glad to see us, if we're boring them, if they're tired, sad, happy, or in a hurry. Facial expressions generally reveal a person's emotions and, to an extent, their thoughts.* This is the basis of social interactions, not only to assess if someone is happy or upset about a comment we might make, but also even for something as trivial as interpreting our interlocutor's gestures, to know when to speak and when to listen in everyday conversations.

The inability to read other people's facial expressions, and, consequently, socially relate to them, is one of the traits that defines autism. People with autism tend to avoid looking at someone's face or eyes. And the eyes are precisely what relay the most information regarding a person's inner state. An autistic person finds it hard to interpret the mood expressed by the eyes illustrated below, while for the average person, these eyes are clearly expressing a state of sadness.**



What mood are these eyes expressing? Is it sarcasm, seriousness, suspicion, or discouragement? (Adapted from Baron-Cohen, *The Psychologist*, 2008.)

Reading the mind, like in *Until the End of the World* and many other sci-fi movies, is no longer a pipe dream; it is a reality that science has brought to light in recent years, be it by interpreting brain signals obtained with fMRI or those produced by neurons recorded with intracranial electrodes. And, as just mentioned, we know that to some extent we read other people's minds by interpreting their unconscious gestures. A mentalist or just a professional poker player meticulously studies these gestures to try to foresee what a person is thinking. This discussion has a lot in common with our previous foray into the theory of mind, the ability to understand someone else's point of view and interpret their thoughts. It is not that we are overwhelmed by curiosity, or wanting to poke around in other people's private thoughts. On the contrary, we seek to understand what those with whom we interact are feeling, because this is the foundation of our social interactions.

---

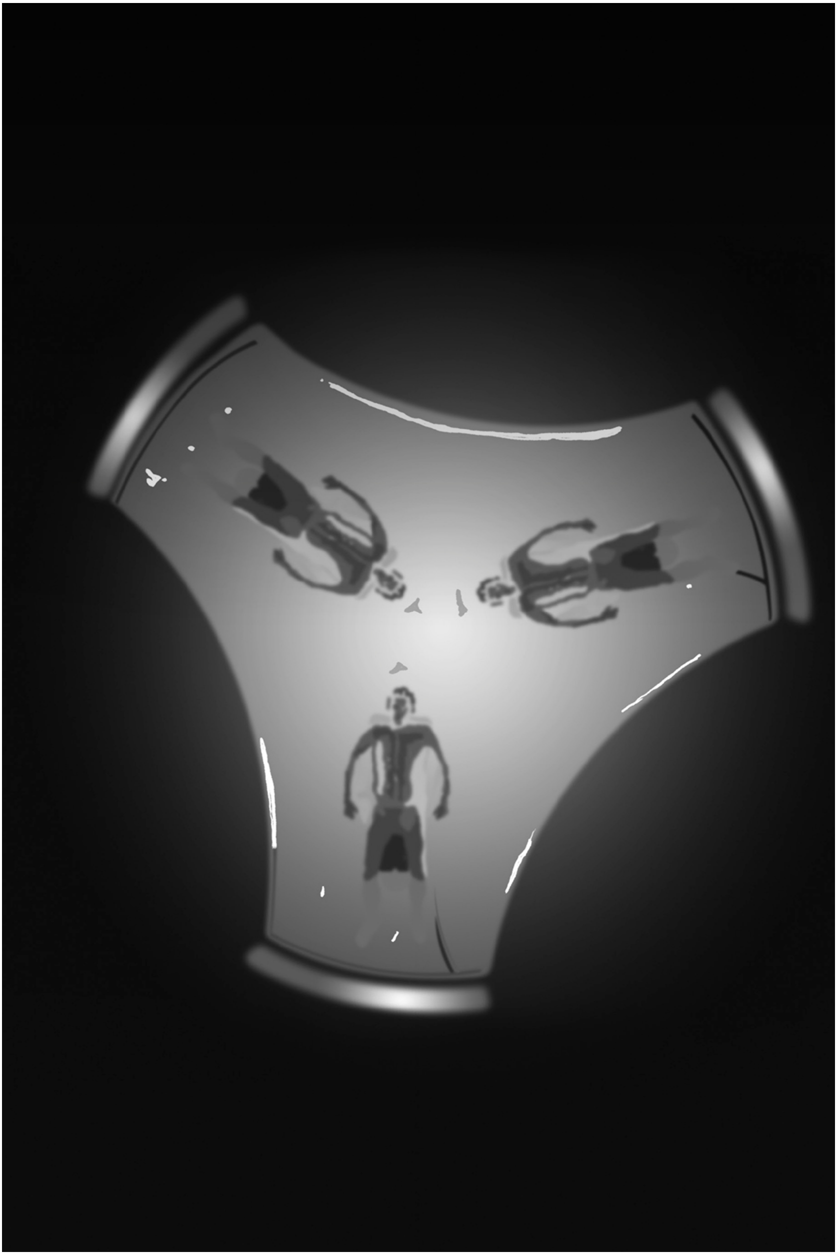
* The pioneer in studying how facial expressions give away emotions is American psychologist Paul Ekman of the University of California, San Francisco.

** The interpretation of the mood expressed by these types of looks is specifically based on a test to evaluate autism. For more information, see Simon Baron-Cohen et al., "Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome," *Journal of Child Psychology and Psychiatry* 38, no. 7 (1997): 813–822. For a more general description, see Simon Baron-Cohen, "Theories of the Autistic Mind," *The Psychologist* 21, no. 2 (2008) 112–116.

CHAPTER 6

MINORITY REPORT

Free Will



In the previous chapter, we explored the possibility of reading someone's mind. We saw how recent scientific studies show that we can effectively ascertain, based on cerebral activity, what a person is seeing or thinking. But can we also predict future actions? Could we, for example, foresee if someone is going to commit a crime? If so, what should we do?

Minority Report was considered one of the best movies of 2002. Directed by Steven Spielberg, the tech noir film is based on a short story by Philip K. Dick.* The action takes place in the year 2054, mainly in Washington, DC, where a division of law enforcement known as *PreCrime* has managed to literally reduce crime rates to zero by predicting—and avoiding—crimes that would inevitably happen in the near future. The predictions are based on the visions of *Precogs*, two men and a woman with extraordinary psychic abilities who are kept isolated, floating in a great big pool—a.k.a. the "photon milk bath"—focused on identifying different crimes hours before they happen.

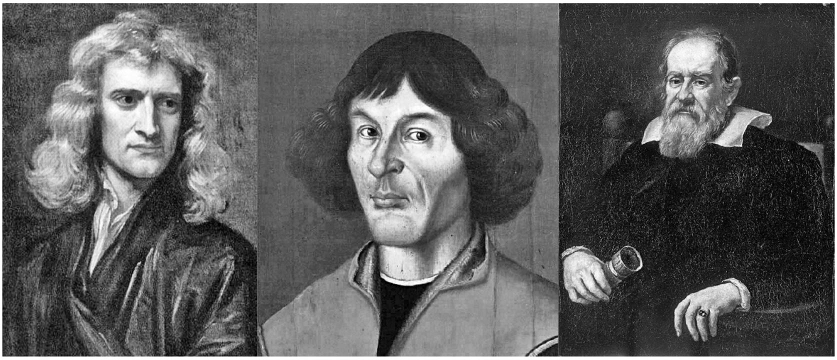
The thriller takes off when one of PreCrime's police chiefs, John Anderton, discovers that the Precogs have predicted he will commit a crime. Fearing the same fate as the potential murderers he usually arrests, Anderton escapes and tries to prove the system's error. Anderton is absolutely certain he is incapable of committing a murderous crime, although he also knows the Precogs' reputation of prediction infallibility and that the die is cast for him. Ultimately, the movie's main theme is the issue of free will, dramatically posing the question of whether we can truly think and act freely if our future is predetermined.

If you saw the movie some years back, like me, you might not remember all the plot details but you may very well recall some of the scenes featuring future technological advances, for example, the one where Anderton, with the finesse of an orchestra conductor, uses the movement of his gloved hands to manipulate data and images on a glass-like, life-size computer screen.

Many of the tech predictions featured in *Minority Report* are already real or not far from becoming a reality. In fact, hand-swipe capability was introduced years ago in smartphones and tablets to manage data and images. Anderton also uses a car that takes him around DC without having to move a finger and, although they are not yet zooming around our streets, self-driving cars are practically here, with Google and Tesla prototypes, among others, in which test models have already traveled thousands of miles. Another prediction, in this case chilling, is the one where Anderton is bombarded by countless personalized ads while walking into the metro, suggesting to him a car brand, having a beer, and other ways to spend his money. Automatic face recognition is already a reality, and the internet is constantly flooding us with personalized ads; by just opening a newspaper's web page, we see Amazon suggestions based on our previous searches. Combining these two advances, it isn't a stretch to think that advertising systems could recognize passersby and offer them different products according to their tastes. *Minority Report* also shows other tech gadgets that are now widespread but back then didn't even exist: smart watches and glasses, like those currently available, a smart home controlled by the owner's voice, and so on.

The number of prediction hits in *Minority Report* is truly striking. It is also no coincidence. Before filming the movie in 1999, Spielberg gathered several science and technology leaders for three days at a California hotel. The goal was to brainstorm and try to catch a glimpse of what the future would be like in fifty years. The movie's revolutionary ideas were the result of that meeting and, furthermore, the gadgets shown were so attractive that countless students, entrepreneurs, and scientists felt compelled to try to replicate them in real life. That exchange among visionaries foreseeing a possible future and entrepreneurs determined to turn such predictions into reality actually make it impossible to foresee how we will be living in a few years—the effects of such interactions go beyond our imagination. We didn't have to wait until 2054; many of *Minority Report*'s ideas are already a reality, less than two decades after its release.

However, this chapter isn't so much about technological advances as it is about the movie's most profound prediction: Will we be able to forecast people's behavior? Will we be in a position to judge someone based on their future acts? This discussion has run amok in philosophy for centuries, since it deals with people's free will and the moral responsibility they have for their actions.



Isaac Newton, Nicolaus Copernicus, and Galileo Galilei: three pillars of the seventeenth-century scientific revolution.

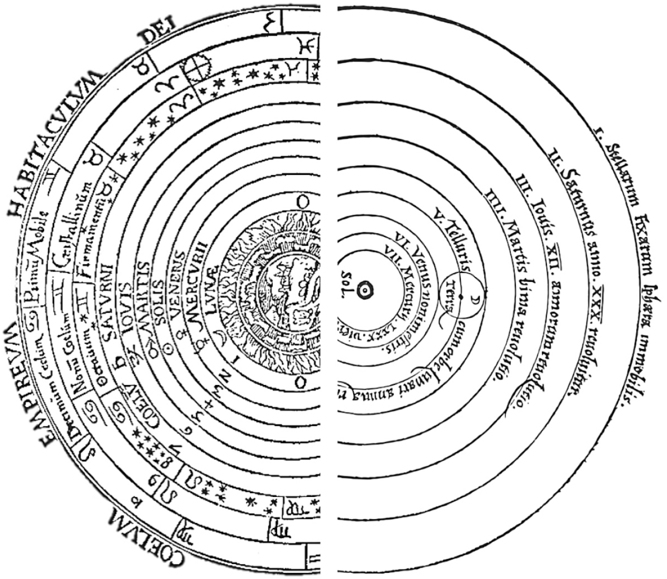
In 1687, Isaac Newton published *Philosophiæ Naturalis Principia Mathematica*, in which he laid the foundation for the modern formulation of classical mechanics. Newton's work marks the highest point of the seventeenth century's scientific revolution, which in physics would overturn Aristotle's (fourth-century BC) postulate about the motion of bodies and Ptolemy's (second-century AD) geocentric concept of the universe.

Aristotle said that to keep a body in motion, constant force must be applied to it. This idea seems reasonable: while a horse draws a carriage, the animal keeps the vehicle moving, and if the horse stops pulling, the carriage stops. A different principle was thought to rule the motion of celestial bodies in the firmament: Ptolemy argued that they orbited around Earth, which also seemed to make a lot of sense since all objects fall toward the center of the Earth, and we see planets and stars circling around it. However, what pushes these celestial objects in their trajectory? Aristotle claimed that Earth was made up of four elements—water, air, earth, and fire—while in the firmament there was a fifth element, ether, which was in constant motion.

Nowadays, you may think it would have been easy enough to reject Aristotle and

Ptolemy's arguments, but that wasn't the case. After all, more than twelve hundred years had to go by for the scientific revolution to come to fruition. Of course, today, everyone knows that Earth is not the center of the universe, but let's pretend we could travel back in time and have a chat with Aristotle and Ptolemy. How would we convince them that Earth is actually the one circling around the sun and not vice versa?

The seventeen-century scientific revolution began to take shape a couple of hundred years before Newton's time, when Nicolaus Copernicus (1473–1543), a Polish astronomer and mathematician, proposed a model to estimate the planets' trajectories, assuming that they orbited around the sun. The main problem with Ptolemy's model was the retrograde motion observed in some planets like Mars: day after day, Mars travels a certain trajectory in the firmament, but once in a while it moves backward.


HABITACVLVM DEI
COELVM EMPIREVM
Primum Mobile
Cristallinum
Firmamentum
Nonu Coelum
Decimum Coelum
SATVRNI
IOVIS
MARTIS
SOLIS
VENERIS
MERCVRII
LVNÆ
I. Stellarum fixarum sphæra immobilis.
II. Saturnus anno. XXX. revolvitur.
III. Iovis. XII. annorum revolutio.
IIII. Martis bima revolutio.
V. Telluris cum orbe lunari annua revolutio
Terris
VI. Venus nono mense.
VII. Mercurij. LXXX. dierum
Sol.

Ptolemy's geocentric model (*left*) and Copernicus's heliocentric model (*right*).

Ptolemy's complex explanation was that the planets revolve around specific points called *epicycles*, which in turn circle Earth. Copernicus's model—avoiding these epicycles when considering that the planets orbit around the sun—was simpler, but it still didn't deliver precise predictions of the planets' motions. Such predictions would only be possible after the German astronomer and mathematician Johannes Kepler (1571–1630) broke away from the idea of Aristotle's perfectly circular paths to postulate the planets' elliptical trajectories around the sun. Kepler's simple and precise calculations themselves constitute a convincing argument for the heliocentric model, but the irrefutable experimental proof would come soon after, with Galileo Galilei's (1564–1642) observations.

After having dramatically improved the resolution capability of the telescope, Galileo discovered, among other things, Jupiter's moons and, consequently, that celestial bodies don't necessarily revolve around Earth. Moreover, Galileo would present the principle of inertia. Contrary to what Aristotle believed, when bodies no longer receive a force, they do not stop, but they continue moving. The carriage does not stop because the horse stops pulling it, but due to friction forces that bring it to a standstill. To use other everyday examples, if we remove our foot from the accelerator, our car continues moving a little while longer until it eventually comes to a stop due to its friction with the air and in its wheels; whereas a craft in outer space continues eternally in motion unless some force, like a nearby planet's gravity pull or its rocket propulsion, changes its trajectory. So, according to Galileo's principle of inertia, we no longer have to imagine ether in motion to understand the movement of the planets; with no friction or other force to stop them in outer space, they continue their (elliptical) trajectory attracted by the sun's gravitational force.

Based on Galileo's observations, Newton later summarized what governs the motion of bodies into three fundamental laws and proposed the *law of gravitation*, from which he deduced the planets' elliptical trajectories postulated by Kepler. As alluded to in chapter two, Newton's great advance was to demonstrate that the principles that rule the mechanics of the bodies on Earth—the fall of an apple—are the same that rule the motion of celestial bodies: the moon's trajectory around the Earth, or the orbit of the Earth around the sun. Thanks to Newton, physics would finally describe the motion of bodies—all bodies—through fundamental laws without needing to turn to strange arguments (the existence of ether in motion, or planets revolving around epicycles) to explain experimental observations.

In this context, born in the dawn of the scientific revolution—more than a century after Copernicus and contemporary to Kepler and Galileo—René Descartes understood that the advance of science should not be limited to past preconceptions, such as those of Aristotle and Ptolemy, and presented the systematic method of doubt as the source of knowledge.

> If our senses betray us . . . and if there are men who err in reasoning . . . I reject as false every reason I previously had as demonstrative . . . I resolved to pretend that everything that had ever entered my mind was no truer than the illusions of my dreams.
>
> —René Descartes, *Discourse on Method* (Part IV)

As we saw in chapter four, in his *Meditations*, Descartes illustrates this idea by creating an evil genius who, using unlimited power, is capable of deceiving us about reality. Such a systematic doubt would lead him to find his most basic truth—*I think, therefore I am*—and initiate the revolution, contemporary to the scientific revolution, that would pave the way to the birth of modern philosophy.

After the publication of Newton's principles, physics reached its pinnacle. And just like Descartes imagined an evil genius capable of deceiving the senses, French mathematician Pierre-Simon Laplace (1749–1827) would conceive, a while later, a similar demon capable of predicting the position of each particle in the universe using Newton's laws of mechanics.

> We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements

> of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.
>
> —Pierre Laplace, *Essai philosophique sur les probabilités*[1]

Laplace's argument is what we know as *determinism*: Each state of the universe is the inevitable and only consequence of previous states. In other words, subject to the laws of the universe, the initial state of an object determines its future states. And this does not only apply to apples and planets in motion, but also to ourselves.

Our brains are is nothing more than the activity of billions of neurons. So, if Laplace's demon knew the state of our neurons in a given moment, he could also predict the brain's future states. Of course, this task goes way beyond our measurement and calculations capabilities. It is impossible to measure the activity of every one of our neurons—and each particle that makes up our bodies, since the brain is not isolated from the body—and even if we had such information, it would also be impossible to calculate future states. But for now, let's set aside practical limitations and continue with the theoretical argument.

In principle, Laplace's demon would know the state of my neurons from my birth to my death. And by knowing the state of my neurons (and my body and everything that interacts with me), he could predict if in a few seconds the neurons in my motor cortex will activate to make me raise my arm, or if those determining the precise movement of my fingers on the keyboard will fire, thus foreseeing the words I write in this moment. The conclusion is devastating: According to determinism, everything I do is predetermined, it is inevitable; it is the logical consequence of the application of the laws of physics, which leave no room for possible alternatives. This is the tremendous blow physics inflicts on philosophy: If everything is predetermined, are we free? Can we choose our future actions? Ultimately, we are asking ourselves if free will really exists.*

Debates about free will are long-standing in philosophy and began way before the scientific revolution. As we saw earlier, in ancient Greece, Democritus proclaimed that in the universe there is nothing more than atoms and the void among them. In this material world, nothing happens randomly; everything has a raison d'être. With these arguments, Democritus aimed to set aside the idea that our actions are determined by the will of the ancient gods, instead maintaining that they have terrestrial causes. However, despite disregarding a supernatural will that explains our behavior, Democritus ended up proclaiming the theory of determinism that would later flourish during the scientific revolution and would end up defying the idea of free will.

Thomas Aquinas (1225–1274), the most prominent medieval scholastic philosopher, and, earlier, Saint Augustine (354–430), saw the need to allude to one of religion's great philosophical conflicts, which is precisely linked to the idea of free will: If there is a God who knows beforehand each one of our actions, are we then free? Could we choose to act in another way? Furthermore, if there's a good and almighty God, why does evil exist? How can a good Christian be free if that person cannot choose to sin?

Thomas Aquinas's and Saint Augustine's answers are the foundation of Christian doctrine. We make decisions freely. God knows the results of our choices, but He doesn't force us to act one way or another. Sin, for these philosophers, equals free will abuse—already present in the biblical account of Adam and Eve's original sin. The good Christian must always do good—if not, they will be a slave to sin—but they have the freedom to choose the best way to do so.

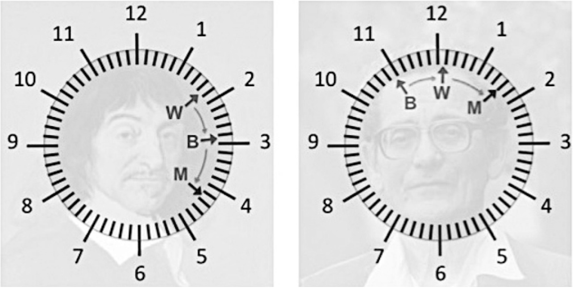
Descartes understood free will as the ability to choose to do or not to do something.

Given the separation he made between mind and matter, Descartes saw determinism and free will as compatible, which in philosophy is known as *compatibilism*. According to this vision, determinism arises from the laws of nature and governs the behavior of bodies and animals. In the material domain, each event is unequivocally predetermined by previous events. However, humans also have a mind with which they make use of their will and freely choose their future. Free will exercised by the mind is the origin of physical events that are unleashed in the material realm.

Therefore, determinism exists, but humans can change the sequence of events by freely making decisions with their minds. For example, if I raise my arm, it is my mind that decided to do so and then (according to Descartes, through the pineal gland) communicated this order to my brain and the different nerves and muscles that execute the motion. The conscious decision to carry out an action precedes its execution. Nothing seems to make more sense. However, in the last few years, neuroscience has completely knocked down Descartes's argument. We have already seen that, nowadays, very few people believe in the separation between mind and brain, but the major blow to Descartes's conception of free will would come through a series of experiments with crushing results.

In 1983, American physiologist Benjamin Libet reported results of what is perhaps one of the most controversial experiments in the history of neuroscience.[2] The experiment's idea is very simple: Participants looked at a screen where a dot traveled a circular trajectory (like a clock's second hand) and, whenever they chose, all they had to do was flex their fingers and then report the second hand's position at the moment they believed they had made the conscious decision to move. Simultaneously, Libet recorded the electromyographic (EMG) activity—the muscle activity measured with electrodes in the forearm—to determine the precise moment when the motion began, and he also recorded the electroencephalographic (EEG) activity to identify the moment the brain was activated preceding the motion. (For many years, it has been known that motion execution is preceded by an EEG activation called *Bereitschaftspotential*, or *readiness potential*.)

In conclusion, Libet could measure three times throughout this experiment: (1) the time the movement was initiated, through the EMG; (2) the time the subject reported being conscious of initiating the movement, which Libet called the *W moment* (for *will*); and (3) the time of the preparation of the movement in the brain, through the EEG.



According to Descartes (*left*), the conscious choice of initiating a motion (W) precedes the brain activation (B), which in turn precedes the muscular motion (M). However, Libet (*right*) demonstrated that the cerebral activation precedes the conscious choice, challenging the idea of free will.

Libet found that the moment of conscious choice (W) happened around two hundred milliseconds before the initiation of the motor action (M). However, what was surprising was that the brain activation (B) not only happened before the motor activation (M), but it also preceded by three hundred milliseconds the moment of conscious choice (W). This discovery wiped out Descartes's idea of free will. The brain activation, then, is not a consequence of the conscious choice of the mind, but rather precedes it. In other words, the conscious decision is a consequence of the unconscious processes in the brain (that later trigger motor activations). If our actions are due to the unconscious activation of neurons —in fact, following determinist rules—apparently there's no choice but to accept that free will doesn't exist.

Most contemporary scientists believe that free will is an illusion or, at the very least, is not what we normally think it is. This vision was already set forth by Baruch Spinoza (1632–1677), who together with Leibniz and Descartes, was one of the seventeenth century's rationalist philosophers. In an excerpt where, with very solid arguments in line with current scientists' viewpoints, he destroys the idea of Cartesian dualism and the separation between mind and matter, Spinoza said this about free will:

> Experience itself, no less clearly than reason, teaches that men believe they are free only because they are conscious of their actions and ignore the causes that determine them . . . Therefore, those who believe they are talking, or quiet, or do anything because of the free decision of their mind, do no more than dream with their eyes open.
>
> —Baruch Spinoza, *Ethics* (Part III, Proposition II)

Libet's experiment not only created great controversies in philosophy but was also highly criticized in neuroscience. The weak point is that the devastating conclusion that free will does not exist was based on estimating point W, the exact moment when the subjects alleged being conscious of deciding to carry out an action.[3] This is obviously hardly precise and also subjective. If I stand in front of a clock waiting to decide to raise my arm, it is very difficult for me to determine the exact moment when I made this decision, and an error of only a fraction of a second could be reason for believing that brain activation precedes a conscious decision. However, recent experiments confirm that Libet's conclusion was correct.

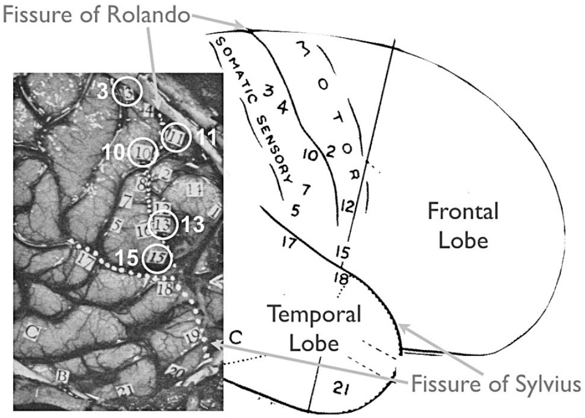
John-Dylan Haynes, a London-educated neuroscientist and currently a professor in Berlin, has devoted his career to using the techniques described in chapter five, analyzing patterns of voxel activations measured with fMRI to predict what subjects are seeing or thinking. Haynes decided to resurrect the idea of Libet's experiment—predicting the cerebral activation that would eventually give way to a motion—but using the activation of voxels instead of EEG signals. The advantage is that, as previously discussed, the responses observed in an EEG are given by the appearance of a readiness potential, which, being generated in the supplementary motor cortex, reflects activations that are relatively late, since this area is one of the last ones involved in generating motion. Therefore, it is probable that previous activations that cannot be observed with an EEG also reflect the decision of carrying out a motion, and much sooner than the five hundred milliseconds reported by Libet.

Haynes's surprising result was that activations seen with fMRI could predict the subjects' motions up to ten seconds before they carried them out. So, although it is true that it is difficult to precisely determine the exact moment when subjects make the conscious decision to carry out a motion (which makes one doubt Libet's results), in Haynes's case, there is no doubt that unconscious cerebral activations precede by several seconds the decision to carry out such actions.

Various experimental results offer even more proof against the notion we naturally have about free will. In chapter four, we discussed Wilder Penfield and his electric stimulation protocol in the exposed brains of epileptic patients. Penfield's objective was to find the different brain functions and evaluate the extent of a potential surgical resection of the seizure focus that would not affect these functions. After studying thousands of patients with this method, Penfield created a map of the sensory and motor cortexes of the human brain, locating areas that, when activated, generated, for example, movements or a tingle in different parts of the body.

In that context, we also referred to my colleague Itzhak Fried and the experiment in which he provoked a patient to laugh when he stimulated an area of the brain. Furthermore, in another group of patients, Itzhak showed that stimulation could not only generate (like Penfield) specific movements, but also the predisposition to carry them out.[4] Penfield's and Fried's results show that, far from having a mind that creates conscious decisions, the determination to carry out a motion, or even the impulse to do so, can be generated by electrically stimulating neurons in different areas of the brain.

Surgery isn't needed to stimulate the brain. It can also be done with noninvasive methods, through magnetic pulses applied on the scalp. This is known as *transcranial magnetic stimulation* (TMS). One of the experts in this technique is Álvaro Pascual-Leone, a Spanish scientist who has been working at Harvard University for several years. Depending on where it is applied, TMS has surprising effects, like stopping speech, creating tingling sensations, causing involuntary motions, and so forth. In a remarkable study, Pascual-Leone and his colleagues asked a group of subjects to extend their index fingers, either on their right or left hand, when they chose to do so.[5] So far, this is no more than refining Libet's experiment: The subjects not only chose when to extend a finger, but also what finger to extend. However, what's interesting is that every so often the researchers applied TMS, causing the extension of one finger or another, depending on where they stimulated. Using this technique, they effectively managed to manipulate the subjects' motion decisions, yet, surprisingly, the subjects were unaware of these manipulations; they reported having made the decision freely, without noticing that, in fact, the motion was no more than the mechanical consequence of the applied stimulation.



Points stimulated by Penfield in a patient (*left*) and the outline of the location of these points in the brain (not all points are in the outline). The fissure of Rolando separates the sensory and motor cortexes, and the Sylvian fissure separates the frontal and temporal lobes. When stimulating points 3, 10, and 15 in the sensory cortex, the patient reported having a tingle in his left thumb, a feeling in the lower left lip, and a numbness in the tongue, respectively. When stimulating points 13 and 11, in the motor cortex, Penfield generated movement in the left side of the face (point 13) and caused the patient to open his mouth and repeatedly pronounce the vowel *A* (point 11).*

From these studies, it is clear that the apparent freedom we have to make decisions can be manipulated without us even noticing it. But the illusion of free will is crucial because, regardless of how fictitious or pliable it may be, the lack of this conscious feeling that we have of being in control of our actions can create devastating pathologies.

In the story "Throat of a Black Kitten" (from the book *Octaedro*, published in Argentina in 1974), Julio Cortázar narrates an atypical meeting in a subway car in Paris. Lucho, a hopelessly romantic Argentine, is approached by a woman who, at first discreetly and then much more openly, begins to caress his hands, while he is holding onto the car's handgrip. Surprisingly, the woman's black-gloved hands are acting on their own, without the conscious approval of their owner, who unsuccessfully tries to apologize to Lucho. This behavior leads to a bizarre situation and, as crazy as it may seem, similar behaviors have been described in patients who have no control over their hands. This is what is known as the *alien hand syndrome*.

Peter Sellers did a hilarious portrayal of this condition in Stanley Kubrick's *Dr. Strangelove*, released in 1964, interpreting a crazy scientist in a wheelchair who cannot control the actions of his right arm. In one of the movie's funniest moments, the rebellious arm is seized and punished by the other arm, but finally frees itself to do a Nazi salute, punch Dr. Strangelove, and then try to strangle him.[6]

Alien hand syndrome is usually caused by a lesion in the corpus callosum—the group of nerves that connects both brain hemispheres—and the frontal cortex. One patient, for example, would be woken up by his left hand, which tried to strangle him while he slept, and during the day unbuttoned his clothes or fought with his other hand when trying to answer the phone. It seems unreal, but, in fact, the man said it was as if "someone on the moon was controlling his hand." In another case, the rebellious hand grabbed a pencil and began to scribble, but when the subject was made aware of what his hand was doing, he was surprised and immediately restrained the rebellious member with his other hand. According to the patient, it was the hand and not him that was doing these things.[7]

So, what can we glean from the cases and experiments described above? The brain's unconscious activity is what initiates action; a short time later we are aware of the action we are going to execute and, ignoring the unconscious processes that preceded it, we attribute the intention (or cause) of doing this action to the fact that we thought about it. This is precisely what creates the illusion of free will—thinking that the conscious intention is what causes our actions—and, as shown by the alien hand syndrome examples, it is truly disturbing when this illusion does not work and unconscious processes directly dictate the functions of a hand that apparently does not respond to our will.*

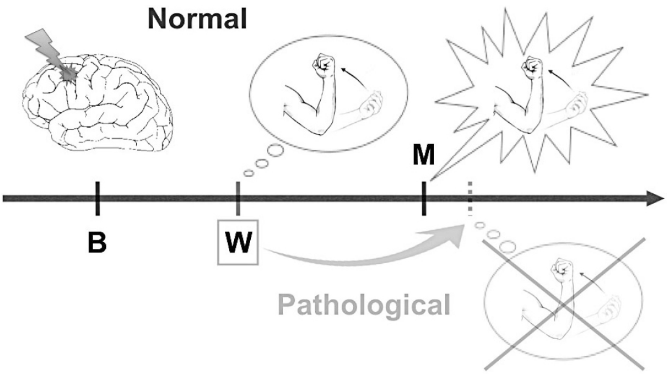
The feeling of not being in control is even more dramatic in patients with schizophrenia. They not only feel they have no control over their actions, but they sometimes also feel they can't control their emotions or thoughts. It is as if they were being handled by external agents.

A popular portrayal of this pathology is given by the character Sméagol (a.k.a. Gollum) in *The Lord of the Rings* film series, or Russell Crowe in his interpretation of famous economist John Nash in *A Beautiful Mind* from 2001. Sméagol's or Nash's thoughts are co-opted by nonexistent subjects and, in some cases, patients with schizophrenia even say that their minds are controlled by aliens; after all, who could have such power to control another person's thoughts at a distance?

However, neuroscience offers a simpler explanation:[8] What is likely failing is the conscious reflection that these individuals have on their own actions, that is, the illusion of free will. In other words, if the process of being conscious of actions we are about to carry out fails, then these actions become unexpected and seem to be responding to external agents. Yet, of course, these actions—be it moving an arm or the words we thread together in a thought—are not generated by others but by the unconscious processes in our brain. It is normal for patients with schizophrenia to report hearing strange voices, but this is nothing other than their own voice, which they can't predict because they are not conscious of what they are going to say before doing so.[9] Let's put ourselves in their shoes for a moment and try to imagine speaking but without having a clue about what we are going to say, surprising ourselves with the words coming out of our mouths. In this situation, it wouldn't be strange to imagine that someone has taken control of our thoughts.

One of the reasons the illusion of free will fails is due to delays in the conscious perception we have of things. If the apparent intention to carry out an action (Libet's W moment) does not precede its execution, the illusion of free will is diluted: Happening to notice that we move before having the feeling that we had planned the movement makes the action seem like it is not responding to our will.[10]

An elegant demonstration of how temporary delays influence the perception we have of initiating our actions was given via an experiment where subjects operated a robot that tickled them in the palm of their hands with a soft-tipped rod.[11] No matter how much we try, we can't tickle ourselves; it is impossible, there's no way. But if someone else is handling the action, even though that person is carrying out the same actions as our own, we hit the roof. The big difference is that when someone else carries out the movements, we have no control; we can't predict what's going to happen. What makes this experiment interesting is that the rod tickling the subjects on the palm of their left hand reproduced their movements with a lever they operated with their right hand. When the movement's replica was instant, the tickling feeling was minimal; however, when a slight delay was introduced in the movement's reproduction, even if it was a split second (which the subjects didn't even notice), the tickling feeling was stronger. This proves that delays in the perception of an action lead us to believe the action is being carried out by someone else.



Unconscious brain activations (B) precede the conscious reflex of, for example, the act of bending an arm (W), which in turn precedes the execution of the motor action (M). Delays in the processing of this conscious reflex (*below*) dilute the illusion of controlling motion (or speech or thoughts), which gives way to pathological behaviors, such as in the case of patients with schizophrenia.

According to Penfield, Fried, and Pascual-Leone's experiments, we have seen that decisions can be manipulated through electric or magnetic stimulation of groups of neurons without the subjects being aware of the manipulations taking place. External stimulations lead to reactions that were somehow "implanted" in the subjects. However, sometimes these involuntary reactions are caused by internal and pathological mechanisms, which are equally uncontrollable.

When I worked as a researcher at Caltech, I met a student with Tourette syndrome.* This condition causes involuntary and uncontrollable reflexes ranging from tics to uttering insults, spasmodic movements (or spitting in the most embarrassing situations, such as in the case of this student). This student once spat on a mutual friend's car window, and in a knee-jerk reaction, he slapped the window to wipe his saliva but ended up breaking the glass. Of course, my friend was far from pleased with his broken window, but what could he say? Not being able to control his actions wasn't the fault of the student.

In Roman law, a crime is punishable if there is both an *actus reus*, meaning a criminal act, and a *mens rea*, a guilty mind or, in other words, the intention to carry out a crime. In the case of the student with Tourette syndrome, one cannot condemn him for his involuntary reactions because there is no *mens rea*; he lacks full control of his actions and cannot avoid spitting, insulting, or breaking a window with his hand. Likewise, one cannot condemn a person with alien hand syndrome for the incontrollable actions of their hand.

Schizophrenia cases are much more delicate since someone with this pathology can, for example, commit a crime apparently forced by an external agent who took over their actions. In fact, if a criminal is proven to have schizophrenia, nowadays they likely will not end up in jail but in a psychiatric hospital. Roman law, in principle, also doesn't punish someone who commits a crime while sleepwalking or while under someone else's influence, for example, through hypnotic suggestions (like in the 1962 movie *The Manchurian Candidate*).

Every one of these cases lack *mens rea*, intent, the conscious process of wanting to commit a crime. These individuals do not exercise their will. But here's the dilemma: How can we even think of judging someone if we argue that free will does not exist? We saw that conscious decisions are the result of unconscious processes in the brain that follow the laws of determinism and are, therefore, inevitable.

Kant argued that without free will there can be no morality. Centuries later, a study showed how the notion of free will affects moral decisions. A group of subjects was asked to read a text on determinism and the nonexistence of free will, while another group was given a neutral text. Both groups took a test, and the researchers proved that the first group showed a higher tendency to cheat. The researchers concluded that, when doubting the existence of free will, the subjects tended to undermine their own feeling of being a person responsible for their acts, which probably gave them the excuse to behave as they desired.[12]

In a conference in Porto, I had the great pleasure of meeting Jim Fallon, an expert on psychopathic killers. And I'm not referring to people who may have killed out of vengeance or following an outburst of rage, even in the most brutal and gruesome way. I'm referring to those like the Hannibal Lecter character, famously played by Anthony Hopkins, who kill in cold blood, almost naturally, with absolutely no emotion or regret.

What is failing in these people? After studying dozens of cases, Jim found a common pattern in their brains due to a deficiency in what is known as *the limbic system*, the one that regulates emotions. It makes sense: The failure in the emotional system subjugates the moral rejection we naturally have when even considering the possibility of killing a person. The variation in the limbic system was so noticeable that Jim could differentiate without a doubt the image of a psychopath's brain from someone else's. But one day, while analyzing an image of his own brain, he found he had the same alteration.[13] University of California, Irvine professor, father, and husband for more than forty years, Jim knew he wasn't a psychopath like the people he researched, but where was the difference?

Jim argued that psychopathic killers not only have a malformation in the limbic system, but also a history of child abuse. That's the explosive cocktail. A person like Jim could kill someone with no sense of guilt, but doing it would never cross their mind. On the other hand, a person with a traumatic childhood could accumulate a level of aggression so great that it would lead them to consider a murder, but they wouldn't act on the impulse due to the repression caused by the limbic system—which translates to empathy and an emotional connection with the other person. However, someone who has both a deficiency in the limbic system and a traumatic childhood could tend to seek out reparation, and no internal mechanism would dissuade them from committing a crime.

And so we return to the main argument of *Minority Report*: If tests like the one developed by Jim Fallon determine the psychopathic behavior of a person, must we, as a society, take preventive measures to avoid a massacre? Can we condemn someone without *actus reus*, that is, before they commit a crime? To answer this question, we must consider that no matter how strong a person's physiological and social tendency to psychopathy, they still have the ability to choose. Besides the fact, as Fallon argued, that all psychopaths have an alteration in their brains and traumatic childhoods, not all people with these two components will necessarily become psychopaths. There may be millions of people with these characteristics whom we know nothing about because they behave normally. In other words, both conditions described by Fallon make a necessary but not sufficient condition. Michael Gazzaniga, one of cognitive neuroscience's pioneers, explains it clearly:[14] There is no alteration in the brain that leads us to commit a crime. A person can have fewer mechanisms of restraint, but they always have the ability to decide, to freely choose. However, it seems we are contradicting ourselves: How can we speak of the freedom to choose if we have already established that free will does not exist?

Most of the confusion about free will comes from what is understood in that regard. We must differentiate the philosophical and neuroscientific problem from moral and law-related issues. It is true that our behavior is caused by the firing of neurons that follow determinist rules, but no one would think of arguing before a judge that, given the conditions of the universe and the firing of billions of neurons, what a murderer did was inevitable.

The law understands free will as acting in *motu proprio*, without being coerced into committing a crime. For example, a person is not guilty of having signed a false check if they can prove that someone else pointed a gun to their head and forced them to do it. What we have been saying about free will in neuroscience—that our behavior is triggered by the activation of neurons following determinist rules—goes beyond considerations about the legal system and gives rise to a much deeper and more philosophical discussion. But let's get back to this at the end of the chapter and now focus on an even more controversial subject.

Unlike the world proposed by *Minority Report*, we agree that a person cannot be judged for their predisposition to commit a crime. A far more delicate situation is that of those who have already committed a crime and are being considered for parole or have already done their time. A discussion about their release back into society involves two different visions on justice: retributive and utilitarian. *Retributive justice* emphasizes fair punishment—what the person deserves for the crime committed. *Utilitarian justice*, on the other hand, prioritizes society's benefit. At first, neither one is more severe than the other. For retributive justice to occur, a convict must serve up to the very last day of his sentence, not one more, not one less. In a utilitarian scenario, the sentence is more flexible because the convict could be released earlier, if they are no longer a risk to society, but they can also remain in jail beyond their sentence if they still pose a threat. This double-edged blade gives rise to controversial estimates on the risk of recidivism as well as the implementation of preventive policies.

In the United States alone, there are more than two million people in jail. The cost of maintaining so many prisoners is exorbitantly high, and a few years ago the possibility of releasing those who had committed misdemeanors and didn't need to be incarcerated was considered; for example, their sentences could be commuted to doing community service. In short, social and economic factors led to the implementation of utilitarian justice for misdemeanors. But how could we determine who should be released and who should not? To address these issues, algorithms were developed to establish a subject's recidivism risk factor, reviewing their criminal record and answers to a questionnaire with more than one hundred questions. Yet what began as an apparently reasonable decision (after all, why spend so much money on prisoners that don't need to be in jail?) has become a monster-like dilemma challenging the US legal system.

The problem is that this type of algorithm is already used in several states, and not just to commute misdemeanor sentences but also to suggest decisions in every type of crime and during the whole legal process: whether a person should be incarcerated or released, the bail amount, if the person should be released on parole, and even the severity of a sentence. The most widely used algorithm, called *COMPAS*, estimates a risk factor that establishes the probability of someone committing a crime again, and this information is given to judges before they declare a sentence. The first big inconvenience is that this algorithm implicitly discriminates: A single, unemployed, twenty-year-old man will have a much higher risk factor than a married, middle-aged woman with children. So, for the exact same crime, the young man will have a higher risk factor and eventually receive a harder sentence. This evidently violates the idea of retributive justice: giving a fair sentence based only on the crime committed. But discrimination is even more rampant because, among the elements used to estimate the risk factor, the social situation plays a pivotal role. A person who lives in a poor neighborhood, with a precarious social environment, will have a much higher risk factor than someone with the same criminal record but from a high-class neighborhood; a person with family members who have committed crimes will have a higher risk factor than someone who doesn't have such relatives. So, people are at least partially judged based on other people's actions, which is very explicit in some of the questionnaire's questions: Have your parents ever been in prison? How many of your friends or acquaintances use illegal drugs?

Examples abound; a black youth who only committed a petty theft was considered twice as dangerous as a fifty-year-old white man with a history of domestic violence, grand larceny, and drug trafficking. Another young black man who resisted arrest without exercising violence was considered three times more dangerous than a white guy who attempted to rob a house.

The errors in these algorithms are considerable: The fifty-year-old man committed another major theft after being released, while the black youth, who apparently was twice as dangerous, didn't commit another crime. The young man who resisted arrest never had another problem with the police, while the white guy was later arrested another three times for drug possession. In COMPAS's defense, it could be said that we are comparing extreme cases, but a detailed analysis of seven thousand cases showed that the algorithm tends to assign higher risk factors to black people than white people. Among those with no repeat offenses, twice as many black people were considered high risk when compared to whites, whereas among those who did have repeat offenses, more than twice as many whites were considered low risk when compared to black people.[15] Initially, we might think these considerations perhaps aren't as important because ultimately the decisions are made by judges. This may be true in some cases, but in others it clearly isn't.

In 2013, a Wisconsin man named Paul Zilly had been arrested for stealing a lawnmower and tools. The prosecutor and defense agreed to a one-year prison sentence, but when the judge saw Zilly's risk-factor report, he immediately revoked the agreement and doubled the sentence. Later, a public defender appealed the sentence and called the algorithm creator as a witness, who stated that the original idea for COMPAS was to reduce sentences and the prison overpopulation, and not to impose or even suggest punishments. After listening to this testimony, the judge reduced the sentence and in his allegation admitted that if he hadn't seen the COMPAS report, he would have given Zilly a six-month sentence—one year tops.

An even more renowned case is that of Eric Loomis, arrested for fleeing from the police in a stolen car. Loomis was sentenced to eight and half years in prison, but what was striking was the fact that the judge explicitly said in his verdict that he partially based his decision on the high-risk factor reported by COMPAS. The big difference is that in this case the judge made explicit what was an open secret. Loomis's lawyers consequently argued that it was unfair to judge him using information provided by an algorithm to which they had no access. After all, it wasn't clear how the COMPAS algorithm assigned the risk factor, and every convict has the right to access information that is used in their sentence. The situation became even more bizarre because the company that provides COMPAS argued that the algorithm is secret and safeguarded by copyright law. So, strangely enough, an algorithm influences a person's sentence and this person has no right to know how those calculations are made because the company's right to profit from sales of the algorithm takes priority. It seems ironic, but Wisconsin's Supreme Court rejected Loomis's appeal, and the sentence was confirmed when the US Supreme Court declined to review the case.

Traveling in the United States, on two or three consecutive flights, all my bags were searched. It wasn't a big deal because I had nothing to hide, but since back then I used to get to the airport right in the nick of time, the unexpected delays added quite a lot of stress to my trips. At some point, I realized that for those flights, the top corner of my boarding pass had an *SSSS* code printed on it. How silly, I thought: If they give me a boarding pass with four *S*s, then I know I'm going to be searched, so if I wanted to bring something illegal, I wouldn't do it. But, on the other hand, I also feared something much more alarming: For some reason that I was unaware of, perhaps I had been put on some blacklist. It is true that sometimes they randomly choose passengers and search them thoroughly, but the fact that it had happened to me two or three times in a row had me worried. Thankfully, it hasn't happened again. It was, in effect, a coincidence (or the fact that the person who checked me in didn't like my face), but it left me thinking: It would be very unfair to end up on a blacklist of this sort without knowing why.

We have referred to algorithms that evaluate the probability of recidivism in people who committed a crime and how such reports are very biased depending on a person's social condition and skin color. It is unfair, but at least we have the foolish comfort (truly foolish and selfish) of knowing that it will never happen to us if we have never done anything wrong. Well, actually, that's not so.

What's so abhorrent in *Minority Report* is that the fictionalized process decides the fate of innocent people who haven't even committed a crime. Astonishingly, something similar is happening in the real world nowadays, and I'm not referring to somebody putting *SSSS* on my boarding pass because I had a stubbly beard. I'm talking about algorithms that are already being used in cities like Chicago and Kansas City, Missouri, to identify subjects who could probably be involved in crimes.* We are talking about people who are judged based on their social condition, their home, their family environment, and even their social network contacts. The argument used to justify the use of these algorithms is that they help prevent crime and, just like high-risk areas can be identified to reinforce police presence, high-risk individuals can also be identified to attempt to deter them—as police effectively do—from committing a crime. But let's put ourselves in those people's shoes. On one hand, having talks and lessons about crime-related consequences seems useful, but, on the other hand, it is unfair to discriminate against these people, put them under a microscope, force them to attend counseling sessions, or have the police keep a watchful eye on them, even though they have done absolutely nothing wrong. Although it is true that they are not condemned beforehand like in *Minority Report*, the typical marginal social situation they live in is aggravated by the discrimination of appearing on a blacklist that, to top it off, is generated by an algorithm whose workings no one knows.

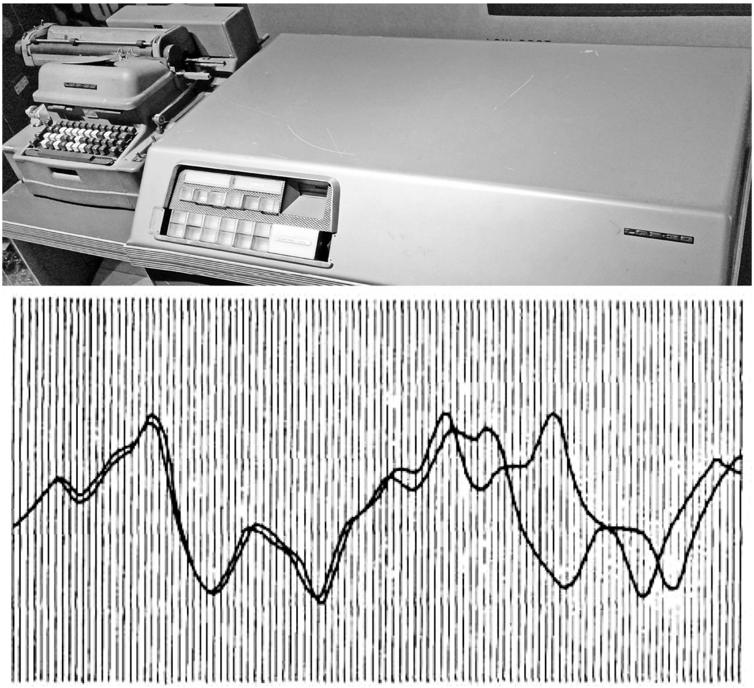
Earlier we argued that science does not offer certainty about the possibility of someone committing a crime. That's the big difference with *Minority Report*: In real life, we're never sure. But what if we were? What if we knew with absolute certainty that a murderer would kill again as soon as they were released? What would we do? Let's put ourselves in the judge's position for a moment and imagine that we are the ones who have to make the decision. The murderer has completed their sentence, and it would be unfair to keep this person past the established time. Even though they were previously guilty, we would be punishing them, like in *Minority Report*, for a crime they haven't yet committed. But we also don't want to release them, knowing that they will kill. There is a clear conflict of rights and interests: the ones protecting the potential victims and society in general, and the ones supporting the convict's rehabilitation. That said, in real life, 100 percent certainty doesn't exist, but what if it was 99, 95 or even 90 percent certain? Where do we draw the line?

This is not just a philosophical problem but quite a real-life situation because the possibility of recidivism of, for example, a rapist or a pedophile is extremely high. The solution implemented by some countries' legislations is to have specialized institutions for reclusion and rehabilitation rather than punishment. So, it is not possible to release a criminal who constitutes a high risk for society, but it is also not possible to continue punishing them beyond the term of their sentence. The situation is perhaps not altogether fair for the convicts, since we continue to deprive them of their freedom. But we also have to consider that they at least had the choice not to commit a crime in the first case, whereas the eventual victim did not choose to be a victim, so their rights should come first.

These issues are undoubtedly controversial, but there's a deeper and likely bleaker discussion pending. We still have to ask ourselves if we are truly free, because if neuroscience leads us to conclude that we aren't, in principle, nothing would seem to make sense. The solution comes from a fortuitous discovery that changed the course of physics.

In the early 1960s, Edward Lorenz, an American mathematician and meteorologist, played Laplace's demon. To do so, he used a computer that took up most of his large office, where he ran mathematical models to try to make meteorological predictions. Just as it was already possible to determine with considerable certainty the future position of celestial bodies, Lorenz's challenge was to develop sophisticated models that could predict the weather as far in advance as possible—a difficult task to say the least, since, despite also being a determinist system, the atmosphere's dynamic is far more complex than the planets' orbits.

The computer, a Royal McBee LGP-30, every so often printed a series of numbers reporting the progress of the simulations, which Lorenz interpreted in order to make his climatological predictions. One day, he decided to continue an earlier simulation. But instead of beginning from scratch, to save time he started in the middle, introducing results obtained from the previous simulation. For example, if he had made a ten-day prediction before, this time he repeated the same simulation but began with the results from day five. In principle, the results for days five through ten should have been the same, since it was the exact same model. But that's not what happened. The results were the same at the beginning, but, as time passed, they were increasingly different.



Royal McBee LGP-30 computer like the one used by Edward Lorenz (*top*) and a series of Lorenz's predictions (bottom), which at first were the same and then began to diverge with the passing days. (Image from Lorenz's 1961 archives.)

Surprised by the result, Lorenz began to investigate the cause of this divergence. It could have been a computer error, which was not usual, or some mistake when introducing the data. However, the computer was working just fine and the data was correct. The difference between the two simulations was due to something that was much subtler: an accident that would change the history of science and would give rise to what today is known as *chaos theory*.*

The problem was that the computer carried out the calculations with six-decimal-digits precision, but only printed three. For example, the computer operated with the number 0.506127, but the printed result was 0.506. When repeating the simulation, Lorenz introduced the results he saw in the printouts to the third decimal point, but the minimal error of not considering the other three decimals made both simulations diverge quickly. And this is exactly the main characteristic of complex and chaotic systems:* Minimal differences in the initial conditions grow exponentially larger. For this reason, chaotic systems are unpredictable.

Let's suppose we have to introduce data in a model to make weather predictions (for example, pressure, temperature, and humidity measured in different places). How many decimal digits should we consider? Chaos theory says that no matter how accurate the precision we use with a chaotic system, however minimal the error, at some point the system will become completely unpredictable.** That is why, nowadays, more than half a century after Lorenz's models, we still can't predict the weather beyond a few days.

The sensitivity to initial conditions is what is popularly known as the *butterfly effect*; according to Lorenz's famous phrase, "the flap of a butterfly's wings in Brazil can influence a tornado in Texas." Clearly, this influence is infinitesimal, and countless other factors would have a much more direct impact on the climate a continent away. But the idea is that no model of a chaotic system, no matter how refined it may be, could consider effects like the flapping wings of a butterfly in Brazil and, sooner or later, this will make predictions useless. Chaos theory radically challenges Laplace and the determinist viewpoint of classical mechanics. The scientific revolution's postulate that, in principle, it would be possible to predict the evolution of the universe is no longer true; there's an unavoidable limitation.

Therefore, though determinism is still valid, complex and chaotic systems are unpredictable. Evidently, one of these systems is the brain. Billions of neurons fire in one way or another to generate our actions, our thoughts, each of the decisions we make throughout our lives. Although it is true that our decisions are inexorably determined by the activity of our neurons and that this activity is, in turn, determined by previous states (and other countless factors), the key is that our behavior is unpredictable. We saw that, using advanced methods, we could predict and even influence if someone is going to move a finger, but regardless of how advanced these technologies may be, we will never know for sure if someone is going to propose to their girlfriend, study architecture, or go to the gym that same afternoon.

Chaos theory is what saves us from the great desolation created by the idea of determinism. Although our actions are predetermined and follow laws of physics, they are unpredictable—there is, and will be, no one who, like in *Minority Report*, can catch a glimpse of our future actions and prejudge us for them. But we are likely still troubled by something that doesn't quite add up: the fact that it seems we have no control, since ultimately our decisions are determined by the firing of our neurons (and not by us). In other words, we don't have, as Descartes stated, an autonomous mind that leads us to use our will to choose one thing over another. But let's delve into this idea. How would an autonomous mind make decisions? If it did nothing more than weigh evidence—that is, putting every reason in favor and against in the balance—however unconscious we may be of these reasons, the decision would still be a determinist process.

The naïve idea we have of free will is that we, in effect, consider evidence, but there's something else that leads us to exercise our will. What is that something else? Maybe personality traits, like the fact that we are honest, emotional, or brave? However, these factors could be explained by our genetic inheritance and our education, also following deterministic rules.

The opposite of determinism is randomness; there are no more options. If determinism doesn't explain our decisions, then we make them randomly, like flipping a coin, or perhaps it is a combination of both. But even admitting, as posited by Aristotle, a random share to free us from determinism—in other words, having the predisposition to choose something, but knowing that by chance we could end up choosing something different—the idea that we, at least partially, choose randomly is also quite far from Descartes's notion of free will.

In conclusion, we are free. We exercise our will. However, free will is not magical or inexplicable, but rather given by the behavior of our neurons. The error lies in thinking that it is our neurons and not us who are in charge of our decisions, since by making such a statement, we once again fall into the Cartesian dualism trap, so rooted in our intuition and so difficult to avoid. Our neurons and us are one and the same thing. If my neurons fire in a specific way to decide what I write in this moment, I am the one who is doing this. The autonomous mind that exercises its will and freely makes decisions does not exist, but what do exist are the neurons and the conscious and unconscious processes—the self—that lead us to do one thing or another. Even if Decartes's naïve idea of free will is an illusion, we are still free.

Ultimately, we are wondering about something that, if we think a bit about it, should be obvious. We really don't need to know about Libet's experiments to conclude that Descartes's idea of free will does not make sense. We also don't need to know much about neuroscience to realize how fragile and malleable our decisions are. All we need is a dose of introspection to glean that in many cases we tend to choose impulsively, and we rationalize our decisions a posteriori. We choose one job over another because, perhaps even unconsciously, we liked our future boss's handshake or appreciated the reclining chair in our new office. Yet we later rationalize and argue that the chosen job is the most appropriate one for some reason or another.

A clever example of the fragility and malleability of our decisions can be drawn from what is known as *choice blindness*. In one such experiment, different people were shown pairs of pictures of unknown women. After seeing each pair, the subjects indicated which woman they found more attractive; then the researcher showed them the chosen picture and asked why they had chosen her. Interestingly enough, using an adaptation of a magic trick, every so often the researcher swapped the chosen picture without letting the subjects notice, and these subjects ended up explaining why they liked the person that they had actually not chosen.[16]

Following this idea, the same group of researchers demonstrated that they could even change people's political tendencies: During a general election period in Sweden, they manipulated (also using a magic trick) the written responses that randomly chosen persons wrote on a questionnaire about political attitudes, and they observed that minutes after filling out the questionnaire, the subjects justified political attitudes that were very different from the ones they had initially manifested in their answers.[17] These experiments reaffirm the fact that we tend to choose quickly and impulsively, and then later construct a scaffolding of arguments to justify those decisions, hence maintaining the idea that we choose based on our reasoning: the illusion of free will.

---

[*] Dick also wrote the story *Blade Runner* was based upon, as discussed in chapter two.

[*] Determinism was later set aside with the emergence of quantum mechanics in the early twentieth century. According to Heisenberg's uncertainty principle—one of the pillars of quantum mechanics—we cannot simultaneously know the exact position and speed of a particle. Therefore, the idea of Laplace's demon is not possible, not due to a practical matter, but because it would violate the laws of physics, which limit our knowledge of the state of the particles that compose the universe. However, beyond this criticism of determinism, the arguments against free will continue to be valid, because the uncertainty in quantum mechanics happens at a microscopic scale, too small to influence the behavior of neurons.

The idea that quantum effects can explain the brain's behaviors was, however, proposed by the acclaimed mathematician Roger Penrose in his book *The Emperor's New Mind* (Oxford, Oxford University Press, 1989). Although this idea, stated without much scientific foundation—almost like saying if we don't understand how the brain works, there must be a quantum effect—was, save for rare exceptions, unanimously rejected by the scientific community. For more information, see Christof Koch and Klaus Hepp, "Quantum Mechanics in the Brain," *Nature* 440, no. 7084 (2006): 611–612.

[*] Figure adapted from: Wilder Penfield, "Some Mechanisms of Consciousness Discovered During Electrical Stimulation of the Brain," *Proceedings of the National Academy of Sciences USA* 44, no. 2 (1958): 51–66.

[*] The consciousness that precedes our actions not only gives us the sense of being in control, but also offers us a veto mechanism to resist carrying out an action before executing it. See, for example: Marcel Brass and Patrick Haggard, "To Do or Not to Do: the Neural Signature of Self-Control," *The Journal of Neuroscience* 27, no. 34 (2007): 9141–9145.

[*] A representation of the consequences of Tourette syndrome is portrayed in the 2008 movie *Front of the Class* (based on the book by the same title), which describes the problems that plagued Brad Cohen, a boy with Tourette's who constantly faced discrimination at school. Cohen is currently a teacher, striving to be the kind of teacher he would've liked to have had.

[*] The use of these methods has been covered in several *New York Times* articles.

[*] Without going into specifics, but to be fair with the attributions, it is worth noting that the origins of chaos theory date back to French mathematician Henri Poincaré, in the early twentieth century. For a thorough and enjoyable introduction to chaos theory, see James Gleick, *Chaos: Making a New Science* (New York: Viking, 1987).

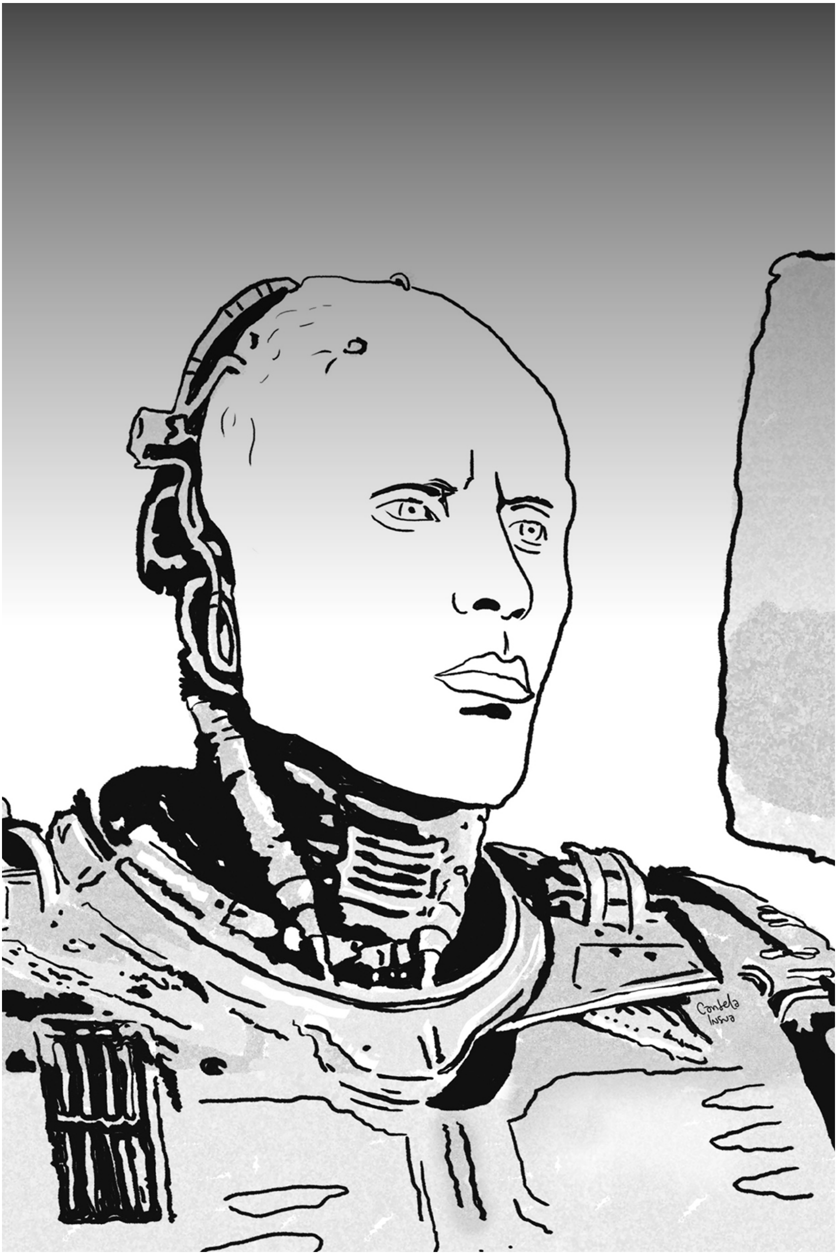
[*] To be clear, mathematically speaking a complex system is not the same as a chaotic one, but in the context of this discussion, it is not worthwhile to delve into these technical distinctions, so I have used the terms interchangeably.

[**] The *Lyapunov exponent* determines the level of sensitivity to the initial conditions or, in other words, how quickly a chaotic system becomes unpredictable.

CHAPTER 7

ROBOCOP

Cyborgs and Identity


Candela
Insua

When I was a kid, I eagerly awaited each episode of *The Six Million Dollar Man*, in which Steve Austin, astronaut and test pilot, had suffered a serious accident when his spacecraft crashed. Series producer Harve Bennett and Richard Anderson, in the role of Oscar Goldman, famously intoned the off-camera voiceovers that launched the plot of the series:

> **Bennett/Narrator:** Steve Austin, astronaut. A man barely alive . . .
> **Anderson/Oscar Goldman:** Gentlemen, we can rebuild him. We have the technology; we have the capability to make the world's first bionic man. Steve Austin will be that man. Better than he was before. Better, stronger, faster.*
>
> —*The Six Million Dollar Man*

Rebuilding Steve Austin cost six million dollars and consisted of replacing an eye with one that he could use to see twenty times farther than the human eye; both legs, which allowed him to run at speeds greater than sixty miles an hour; and his right arm, which was designed to have extraordinary strength. I don't remember specific story lines from those long-ago episodes, but I certainly recall scenes where Austin, played by Lee Majors, runs at the speed of a car, leaps over walls, or watches what happens at a great distance.

The series was a smash hit in the 1970s. It was based on the original *Cyborg* novel by Martin Caidin, and that's what Austin was, a cyborg (cybernetic organism), a human with bionic parts: a hybrid between man and machine. Like many of my age, I grew up watching *The Six Million Dollar Man* as something fascinating and unattainable. But today we are surrounded by cyborgs—people with cochlear implants, artificial organs, or tiny pacemakers—and replacing a person's arm, legs, or eye with synthetic prostheses is no longer far-fetched. A similar idea can be found in the movie *RoboCop*, where police officer Alex Murphy is reconstructed after suffering serious injuries and continues fulfilling his role, but using the extraordinary capabilities provided by his bionic parts.

In the previous chapter we discussed the possibility of anticipating crimes and eventually preventing them. In *RoboCop*, the idea is to reduce the crime rate by using almost unbeatable police officers to face criminals. An interesting aspect of the movie is that it raises the possibility of using robots, instead of cyborgs, for this role. But beyond the amazing advances in artificial intelligence (see chapter one), it is still preferable to have a human in control, since, even if it is a cyborg, its human brain is the one making decisions using common sense and emotions, something machines do not (yet) have. The possible consequences of a program error in a robot carrying weapons and using brute force are terrifying. The beginning of the movie actually shows how a security robot misinterprets an innocent person's intentions and riddles him with bullets—which leads to the decision to use cyborgs.

Putting a human in control imposes important challenges though. It is no longer as easy as building an artificial arm that to some extent replicates a human arm's functionality—

something that has been largely achieved these days. The true complexity lies in driving the mechanical arm via brain activity, that is, reading the brain's intentions of movement and communicating them to the arm or any other bionic limb. This type of development is known as *neural prosthetics*, a booming research field in recent years.[1]

The primary objective of the advances in this field is not to create cyborgs like RoboCop or Steve Austin, but to develop tools with which tetraplegic patients (or those with amputations) could command prosthetic limbs to reach objects, feed themselves, or walk again. That's the main line of research at Richard Andersen's lab,* who, together with Christof Koch, was my mentor at Caltech. I remember being stunned when I first went to his lab and saw firsthand that the possibility of using neural implants paired with prosthetic limbs wasn't far from becoming a reality. The idea of *The Six Million Dollar Man* was no longer a pipe dream.

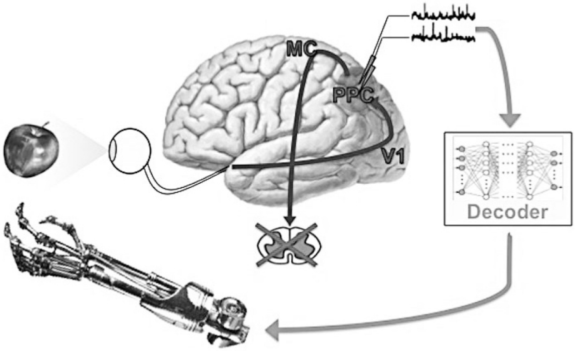


Diagram of a neural prosthesis. To grab, for example, an apple, seeing the object creates activations in the primary visual cortex (V1) that are transmitted to the posterior parietal cortex (PPC), where the movement's intention is generated. These activations are later transmitted to the motor cortex (MC), where the execution of the movement is implemented, and to the arm's muscles through the spinal cord. In patients with a spinal cord injury, the signal doesn't reach the muscles. However, the person's intention to move can be decoded through the recorded neurons in the PPC (or MC), with the information then transmitted to a mechanical arm that carries out the movement.

Richard's most important scientific achievement to date is the discovery of two areas in the posterior parietal cortex that are dedicated to planning the intention of different motions (of the arms and the eyes).* The posterior parietal cortex has a privileged location for this function because it is between the primary visual cortex—where the view of, for example, an apple is coded—and the motor cortex, where the execution of a motion to reach for it is planned, as referenced in the illustration on page 157. This is what inspires the idea of neural prostheses. A patient such as the actor Christopher Reeve—who after portraying Superman tragically ended up confined to a wheelchair after being thrown from a horse and suffering a spinal cord injury*—can see the object he wants to reach and can also plan the movement to reach it. (That is, the intention of the movement, coded in the posterior parietal cortex and the motor cortex, is similar to that of an uninjured person; the problem is the damage in the spinal cord doesn't allow him to communicate this signal to the muscles of his arm to execute it.) Then, the neural prosthesis reads the neural activity via electrodes implanted in the brain and, using decoding algorithms (which were described in previous chapters), predicts the movement the person wants to make and carries it out with a mechanical limb.

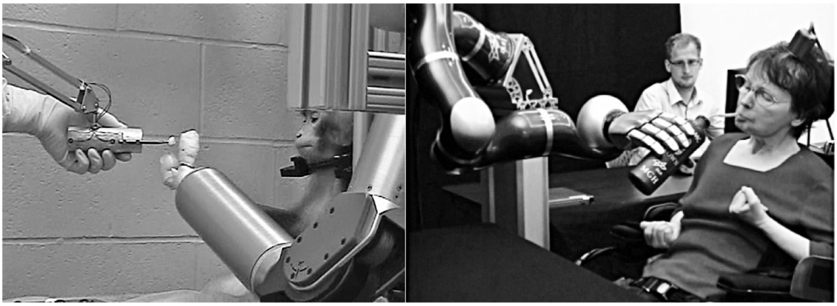
The idea of a decoding algorithm is similar to the one described in chapter five, where we saw that analyzing the recordings of neurons in patients who looked at different pictures allowed us to predict what picture they were seeing. In this case, neurons in the posterior parietal cortex do not respond to Halle Berry or Jennifer Aniston—like those aforementioned neurons—but to the intention of carrying out specific movements. For example, one of these neurons will respond each time the subject plans to move their arm down, another will do so when the person wants to move their arm to the left, and so on. Using the activity of a few dozen neurons, it is then possible to perfectly predict the movements of the arm, as shown by Andersen's experiment with monkeys.[2]

Once we can precisely read the movement's intention, all that is left is to activate the mechanical arm with these signals and *close the loop*: have the subject see the movement of the mechanical arm and be able to control it through the activity of their neurons, correcting eventual errors through visual feedback. The great advantage is that decoding specific details of the movement is not necessary; robotics takes care of that. If the brain sends the instruction to move the arm to a specific location, the mechanical arm's control algorithms (and not the brain) can take care of implementing a fluid movement toward that point, coordinately applying different forces to each articulation. This greatly simplifies the problem because we only have to decode general instructions and not the details on how to implement them (just like we're only aware of wanting to move an arm to a certain place and not of the different forces we must apply to each of the involved muscles to carry out the movement).

One of the most spectacular results in this field was published in 2008 by Andy Schwartz's team at the University of Pittsburgh, when they showed that a monkey (with its own arms gently restrained) was able to continuously control the movement of a mechanical arm through the activity of a few dozen neurons. The way the monkey controlled the mechanical arm was so precise that it could grab pieces of food placed in different spots in front of him and bring the food to his mouth (see left image on page 160).[3] This result not only caused a great commotion among scientists but was also widely covered by the press.[4] The next question was obvious: if Schwartz and his colleagues had achieved such a remarkable advance with monkeys, could a similar technology be implemented with humans?

It didn't take long to find that answer. A few years later, in 2012, the scientific community—and the general public, after impressive widespread coverage by the media—witnessed how similar results were obtained in tetraplegic patients.[5] The right image on the next page shows one of these patients, who managed to manipulate a bottle to drink its contents, marking a milestone in the history of neural prosthetics.

The advances described still differ greatly from Steve Austin or RoboCop, but they constitute *proof of principle*, a credible demonstration of the possibility that a person can control biomechanical prosthetics directly from their brain activity. How far off are we from developing sophisticated cyborgs like those in the movies? Are they still just part of the collective imagination, or will we see these cyborgs walking on the streets in the near future?



Monkey grabbing a marshmallow with a mechanical arm controlled by his brain, which it would later take to his mouth (*left*). Tetraplegic patient handling a mechanical arm through the activity of her neurons to reach for a drink (*right*). (Adapted from Velliste et al., *Nature*, 2008; and Hochberg et al., *Nature*, 2012.)

An important aspect that has delayed substantial progress is strictly ethical and technical, and less so scientific. These types of neural interfaces require a major surgery to implant dozens or hundreds of electrodes in the brain. Here's the looming question: Is such a surgery justified given the risks it implies?

We have seen that the advances in robotics allow us to carry out somewhat complex tasks—like reaching for a drink and bringing it to the mouth—and if we could guarantee patients that with the implanted electrodes they could perform these types of tasks for the rest of their lives, then the surgical intervention would be widely justified.* However, if the surgery only allowed them to command neural prosthetics for a relatively limited amount of time, it wouldn't seem to make sense. (This discussion isn't entirely black-and-white, since even if it didn't make a difference in the implanted patient, the surgery could also be justified if it gave way to advances that would eventually allow future patients to have a better quality of life.)

The problem is that, although it is true that by recording the activity of dozens or hundreds of neurons it is feasible to handle a neural prosthetic, it is still not possible to maintain stable recordings of these neurons over several years. For the first tetraplegic patients implanted with electrodes, recordings were kept only during a constrained time frame. This was obviously disappointing for the patients, because the decay in the quality of the recordings implied that they lost the ability to command the mechanical arm they were previously able to control.

Current results are much more promising, with recordings spanning several years.[6] However, these recordings are not stable yet. In other words, dozens of neurons can be recorded over several years, but the neurons being recorded change daily; some disappear and new ones pop up.* So, how can the decoding algorithm work? Let's say we train the algorithm to consider one neuron's activity that codes for upward movements. Each time this neuron fires, the algorithm interprets that the patient wants to move their arm up. But the algorithm will obviously get confused if sometime later the conditions of the recordings vary, and another neuron appears in this one's place, coding movements toward the left.

The solution to this problem stems, at least partially, from *neuroplasticity*: Neurons adapt their activity to make the system work. Following the above example (and slightly simplifying it so that the idea is clear), if one neuron codes upward movement and it is replaced by another one that codes movement toward the left, this new neuron could adjust its activity (for example, increasing the coding of upward movements and decreasing the one for movements toward the left) to make the arm move coherently. This type of neural adaptation was demonstrated in an experiment where researchers first trained a monkey to move a cursor on a screen (using a decoding algorithm based on the neurons' activity), then disrupted the algorithm by changing the information given by an ad hoc selection of the recorded neurons, and finally they observed that some neurons adapted their responses to compensate for the disruption so that the monkey was able to continue coherently moving the cursor on the screen.[7]

This all means that through neuroplasticity—and the use of adaptive decoding algorithms*—we can deal with the variability in the recordings to make neural prosthetics last several years.** But, if we seriously consider making developments like those in *RoboCop* or *The Six Million Dollar Man* a reality, we still have important issues to address.

The opening illustration of this chapter depicts the moment when Alex Murphy/RoboCop takes his helmet off and sees for the first time what he has become: a cyborg. After examining his reflection, part human and part robot, he asks about Murphy's family, saying, "Murphy had a wife and son." Interestingly, he doesn't ask about his own family, but rather the family of Murphy, in the third person, as if asking about someone else's wife and child.

RoboCop was released in 1987, and in a 2014 remake, a similar scene is even more dramatic. After recovering consciousness, Murphy asks Dr. Dennett Norton,*** the man in charge of the surgery, about the "metallic suit" he is wearing. Norton tells him it is not a suit but part of his body. Then he shows him what's left of his body, and Murphy faces a mirror and sees how, one by one, his robotic members are disassembled: legs, arms, abdomen, thorax. His despair grows with each piece he confirms is no longer his own body, and he appears to feel as if his life is over, as if that person he once was—his own self—had ceased to exist. However, claims the doctor, there was no other way of keeping him alive and, to comfort him, Norton says that Murphy's body may have disappeared, but he is still there. But is the cyborg looking in the mirror still Murphy?

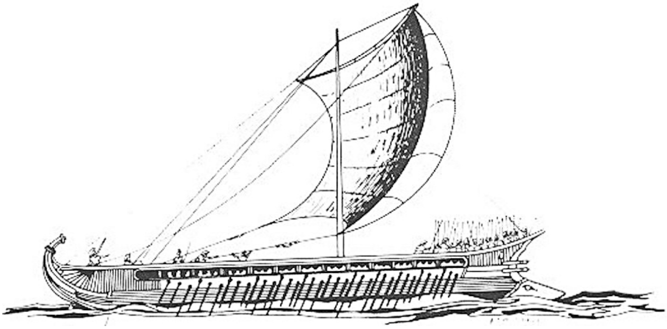


Illustration of the ship of Theseus. After centuries of maintenance, the deteriorated pieces of Theseus's mythological vessel were replaced. But does it continue to be the ship of Theseus?

The scene speaks to the issue of identity, a discussion that goes back to the famous Ship of Theseus paradox described by Plutarch near the beginning of the second century AD.[8] Legend has it that the powerful King Minos of Crete demanded that the Athenians (whom he blamed for the death of his son) pay a yearly blood tribute and send him seven young men and women, who were left in an insurmountable labyrinth to be devoured by a half-man, half-bull creature known as *the Minotaur*. During one of these trips, Theseus, the son of King Aegeus and mythological hero of Athens, accompanies the young group to face the Minotaur. When they reach Crete, Ariadne, the daughter of King Minos, falls in love with Theseus and gives him a ball of yarn so he can mark his path and find his way out of the labyrinth. Theseus slays the Minotaur, escapes the labyrinth, and, after ending the brutal sacrifice of young Athenians, triumphantly returns to Athens in his ship.*

Theseus's ship became a symbol for Athenians who, after having been freed from the unjust subjugation in Crete, every year, and for centuries, sent the ship to the Island of Delos to worship Apollo.* With time, its wooden planks began to crack and had to be replaced with new ones. The monumental question that Plutarch then raises is if that ship, lacking the original planks, continues to be the ship of Theseus.

Plutarch's posit is related to the *sorites paradoxes* (from the Greek: "heap"): If we take one grain from a heap of sand, it will continue to be a heap of sand; but if we repeat this process recursively, taking one grain of sand at a time, we will always say that we continue having a heap of sand (if *N* grains are a heap, *N*-1 grains are also a heap), until eventually we are left with nothing. Therefore, gradual alterations mask the perception of change.**

Returning to the Ship of Theseus paradox, the decision regarding whether it continues to be the same boat or not seems to depend on how quickly the transformation takes place. The new planks, not yet eroded by the passage of time, will look somewhat different, but if every few years only one or two planks are replaced, we would tend to say the restored ship is the same. However, if someone decides to replace a large part of the ship overnight, it would be hard to accept that it continues to be the same vessel.

The same reasoning can be applied to the discussion of identity dramatized in *RoboCop*. If he gradually had a leg replaced, then another, then an arm, and so on, RoboCop might never feel like he is no longer Murphy. But if the transformation was sudden, like in the movie, the person would face an identity crisis. Therefore, the body's continuity—the fact that changes in the body are slow going—is what apparently defines a person's identity. After all, we continue to be the same person despite getting older and having our appearance change as the years go by. These arguments seem obvious and convincing on the surface, but the discussion about what determines identity is indeed much more complex.

The idea that personal identity is connected to the continuity of the body was challenged by John Locke. To illustrate his position, he proposed the famous parable of the prince and the cobbler: One day a prince's mind, with every memory of his past life, is transferred to a cobbler's body. "Who is who?" wondered Locke.[9]* Will the cobbler become prince now that he has the royal son's mind? Locke argued that identity goes with memory because the prince will continue to recognize himself as such even if he is in someone else's body. This view is radically different to what we postulated with Theseus's paradox; in Locke's version, what defines the person is not the continuity of the body, but of memory.

Franz Kafka's novella *Metamorphosis*, published in 1915, offers an even more compelling example. One day, a traveling salesman named Gregor Samsa wakes up and realizes he is slowly becoming an abominable insect. Yet Kafka fluidly continues the story in the first person, and we have no problem assuming that, despite the monstrous deformation of his body, Gregor and the insect continue to be the same. However, our perception of identity is very different if what changes is our memory. We tend to say that someone who has suffered a devastating memory deficit due to advanced-stage Alzheimer's disease "isn't the same anymore." The body hasn't necessarily changed, but the person's identity has been disappearing with the advance of oblivion.

And so, RoboCop will continue to be Murphy while he keeps his memory.* (Although in the movie, his handlers try to erase it so that he will be a more effective soldier and make automatic decisions with no hesitation imparted by his past memories and emotions.) Neuroscience also reveals that, after some time, Murphy will end up feeling that his implanted mechanical parts are part of his body, much like the planks replacing the eroded ones become part of Theseus's boat.

One of the simplest yet most surprising demonstrations in neuroscience is the *rubber hand illusion*:10 A subject places one hand under a table (or behind a divider, out of their own sight) and where their hand would normally rest on the table, there is a rubber hand. Two artist's brushes caress both hands (the real one and the rubber one), and after a while, the coherence between what the person feels on their real hand and what they see on the rubber one creates the illusion that the rubber hand is part of their body. The illusion is so strong that the rubber hand doesn't even have to look like the real one (a rubber glove filled with cotton will do), and many times the experiment ends with a hammer suddenly coming down on the rubber hand, which makes the subject jump in fear because despite not feeling anything, in that moment they can't avoid believing they just got their hand slammed down with the hammer.

For the rubber hand illusion to work, the tactile stimulation on the real hand must sync with the visual stimulus applied on the rubber hand. The illusion immediately disappears if different fingers are stimulated or if both hands are not stimulated simultaneously, because the expectations created by both senses do not correspond to each other. In the previous chapter we described the contrary effect—where people's hands don't respond according to their thoughts, and they experience alien hand syndrome. The degree to which our hand behaves according to our expectations is what gives us the feeling that it belongs to us, that it is part of our body.

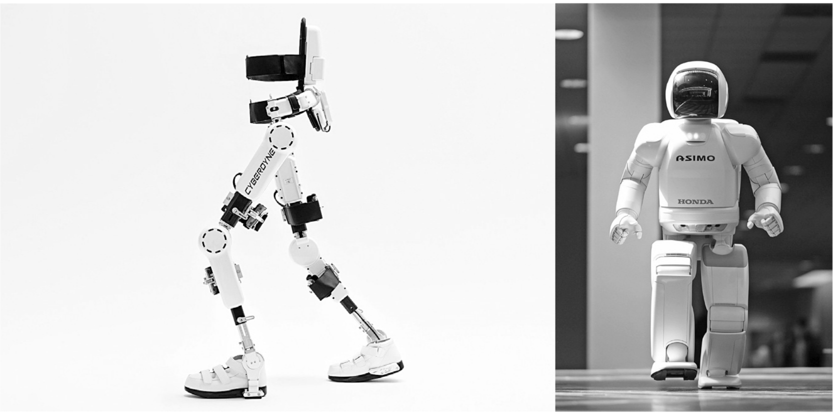
Rubber hand illusion: After brushes are used to caress the same fingers in the real hand (*right*) and the rubber one (*left*) for some time, the subject believes the rubber hand is part of their body.*

An amazing variation of this experiment was developed by Swedish neuroscientist Henrik Ehrsson, who made his subjects feel they were leaving their own bodies, or that their bodies were transferred to a doll's body. To do this, Ehrsson showed how he applied visual stimuli on a doll or any other avatar while simultaneously (and out of sight) applying the same (tactile) stimuli on the subjects. The result is that the coherence between the stimuli seen and felt made the subjects perceive that their bodies were physically the ones Ehrsson was stimulating, even if it was the body of a doll. Besides rationally knowing that their bodies were still the same, the illusion was inevitable.[11] (Using this principle, virtual reality games could be developed in which the player feels tactile sensations that coincide with what is happening onscreen, making them feel they really are the game's avatar as they sense its vicissitudes in the flesh.)

Going back to *RoboCop*, as long as his mechanical members move according to his will and expectations—in the context of the last chapter's discussion, creating the illusion of free will—he will feel they are his own. The same applies to tetraplegic patients. If the implanted members follow their brain's instructions (decoded through their own recorded neurons), it is likely that the patients would end up feeling that those prosthetic members are part of their bodies. In other words, a synthetic arm that does absolutely nothing, no matter how similar in appearance it might be to an original arm, will never cease to be an external prosthetic, whereas an arm that moderately reacts to the patient's will, regardless of how different it may look, would become part of their body.

So far, so good, but we must confront yet another big dilemma. Would we deem it acceptable to amputate a tetraplegic patient's limb in order to replace it with a mechanical prosthesis? We have no problem agreeing to this in the context of a sci-fi movie, but would we do it in real life? Let's put ourselves in the patient's place. Their paralyzed arm is slack and useless, but it is still their arm and, despite feeling absolutely nothing, they wouldn't want to see it hurt and even less so amputated. Although we have addressed the identity dilemma and can assume the patient would end up feeling that the new limb is part of their body, the decision is not clear. However, recent advances offer less dramatic solutions.

For the inauguration of the 2014 World Cup in Brazil, Miguel Nicolelis, one of neuroprosthetics' pioneers and a big soccer fan,* promised something seemingly impossible: that a disabled patient, in front of the entire stadium and hundreds of millions of people following the live broadcast on TV, would get up from his wheelchair, walk a few steps, and take the initial kickoff using an exoskeleton—an external robotic support attached to his limbs—controlled by his brain. As it turned out, the promise was diluted, and the network showed a three-second shot of this person, held up on both sides, barely moving his right leg a few inches to "kick" the ball. However, the disappointment of what would have been an extraordinary demonstration is nothing compared to the incredible advances in recent years. In fact, nowadays, exoskeletons—such as Cyberdyne company's HAL model—are already being sold to help execute heavy tasks and also assist with the rehabilitation of stroke patients. For models like HAL to be used by tetraplegic patients, limbs would have to be controlled directly by the brain. As we saw earlier, this is already within current science's reach.



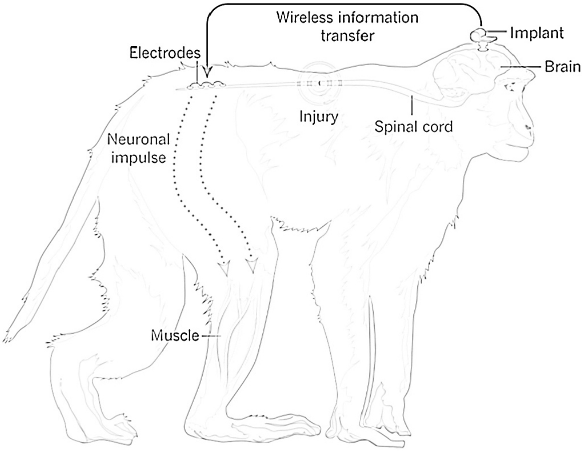
Cyberdyne's HAL (Hybrid Assistive Limb) exoskeleton (*left*) and Honda's ASIMO robot (*right*).

Walking involves the coordinated activation of different muscles and joints to maintain balance and a smooth motion. Regardless of the complexity involved in coordinating these movements, they do not need to be directly controlled by the brain. Like the case of the mechanical arm, robotics takes care of this. Robots like Honda's ASIMO can already walk, run, jump, and move with incredible precision and smoothness. The instructions sent by a person's brain to command prosthetic members can, then, be very general: sit, walk, bend, run, and so on. The application of these instructions is an issue of robotics that has already been practically resolved (which doesn't mean it can't be improved).

Again, as with the mechanical arm, not having to control the details of different movements greatly simplifies the problem, and the use of exoskeletons frees us from having to wonder if we should or shouldn't amputate a patient's disabled limb to replace it with a mechanical prosthesis. But if we accept letting go of the notion of handing patients extraordinary abilities—like Steve Austin's strength and speed, which isn't giving much up since that wasn't the main objective anyway and, furthermore, patients may not want to carry exoskeletons like RoboCop—then there's a much more practical and natural solution.

Grégoire Courtine is one of many physicists devoted to neuroscience who has been exploring the possibility of recovering locomotion through spinal cord electric stimulation in his lab at the University of Lausanne. The big difference with the work we previously discussed is that, in this case, the idea is to send brain signals not to external robotic limbs or exoskeletons, but to the patient's own nerves and muscles. In one study, Courtine and his team injured a monkey's spinal cord, completely paralyzing his right leg. Then they recorded the activity in the motor cortex responsible for the limb's movement, decoding the information and transmitting it to the muscles through spinal cord electric stimulation (in a spot right below the injury). In other words, when the transmission of the neural signal through the spinal cord was blocked, they read the information above the injury (in the cerebral cortex) and transmitted it to a spot below the injury (in the spinal cord), thus managing to restore the monkey's walk.[12] The demonstration was amazing: With no stimulation, the monkey limped and dragged his leg on a treadmill, but as soon as it was stimulated, it started moving its leg, acquiring a much smoother walk.

The development for humans is no less astonishing. By constantly stimulating the spinal cord, it is already possible for tetraplegic patients to move their limbs and even leave their wheelchairs and walk with a walker, rekindling Miguel Nicolelis's promise at Brazil's World Cup.*



The concept of a natural neuroprosthesis developed in monkeys. The signal measured through implants in the motor cortex is transmitted to electrodes that stimulate the spinal cord (under the lesion), which in turn produces the activation of the paralyzed leg's muscles. (Adapted from Jackson, 2016.)[13]

The solution implemented by Grégoire Courtine and his colleagues is very natural because it reuses the patient's own muscles to carry out the movements. The disadvantage of not using mechanical prosthetics or exoskeletons is that we no longer can rely on advances in robotics to resolve the complex muscular coordination that allows us to walk. However, we expect that the previously described neuroplasticity, along with the use of advanced decoding algorithms, will help make such movement possible. In other words, even though the connection between brain and muscles varies (the same connection will never be restored to what it was), the combination of the use of sophisticated algorithms and the adaptive capability of neurons will help solve the problem.

A similar principle was also used to restore movement in a monkey's paralyzed arm, but in that case, they directly stimulated the arm's muscles (instead of the spinal cord) through neural activity recorded in the motor cortex.[14] Like the experiments mentioned above, this result was soon followed by similar experiments in humans, where signals recorded in the brain were used to stimulate muscles in the hand, allowing a patient with a paralyzed arm to move her fingers, manipulate a bottle, and serve its contents in a glass.[15] But grabbing a bottle of water is clearly not the same as picking up, for example, an egg. If we grab the egg with the same pressure we use to grab the bottle, we will break it. Similarly, if we grab a bottle with the same pressure with which we grab an egg, we will drop it. So, what is needed is the ability to control the pressure exercised by the hand to manipulate different objects.

We're so used to adjusting the pressure we use to pick up and manipulate various objects that we are hardly aware of it. We mainly regulate the force we use based on tactile feedback (processed by the somatosensory cortex) received through the pressure of our fingers on the object at hand; it suffices to try to light a match while wearing gardening gloves to understand the importance of the fingers' tactile feedback. One step in this direction was achieved by Miguel Nicolelis when he managed to make a monkey not only control the movements of a cursor on a screen through the activity of its motor cortex, but also to differentiate three virtual objects associated with three different textures based on stimulation applied to the somatosensory cortex, similar to the stimulation received when touching an object.[16]

So far, we have seen that recent advances will soon allow restoration of limb function in tetraplegic patients, be it using their own members or replacing them with synthetic prostheses. But how far are we now from replacing sensory organs? It is one thing to command a leg or arm and quite another to create sensory prostheses that transmit information to the brain to interpret what happens around us—everything we see, feel, and hear.

The Six Million Dollar Man's success led to a spinoff series, *The Bionic Woman*, featuring a female character who had both legs, her right arm, and right ear replaced after a parachuting accident. The actual surgery used to restore hearing, through a cochlear implant, is precisely what constitutes the most successful example of neural prostheses.

In external space, there is mainly silence. Sound, as we know it,[*] is created by variations of air pressure, and it exists in an infinitesimal fraction of the universe; it is a practically exclusive property of our planet Earth and of other planets that could host an atmosphere. Even more, sound itself, be it the voice of a friend or a blast of thunder, doesn't even exist in the atmosphere. It is a construction by the brain through the vibration of small filaments in the cochlea (the inner ear) that transform variations of air pressure into nerve impulses. In cases of profound deafness due to problems in the inner ear, the idea is to record sounds through a microphone and convert them into nerve impulses via electric stimulation with electrodes implanted in the cochlea.[**]

Since the show was first featured in the 1970s, the hearing problem of *The Bionic Woman* has already been largely resolved and, today, hundreds of thousands of patients can hear thanks to cochlear implants. However, it is much harder to get someone like Steve Austin to recover his sight.

Following the success of cochlear implants, entrepreneurs and scientists faced the challenge of developing retinal implants. The principle is the same, but instead of capturing sound, the retinal implant captures rays of light that reach the eyes through a sheet with photoreceptors placed on the retina (or using a video camera mounted on special eyeglasses) and converts them into electric currents applied through microelectrodes that stimulate the optic nerve.

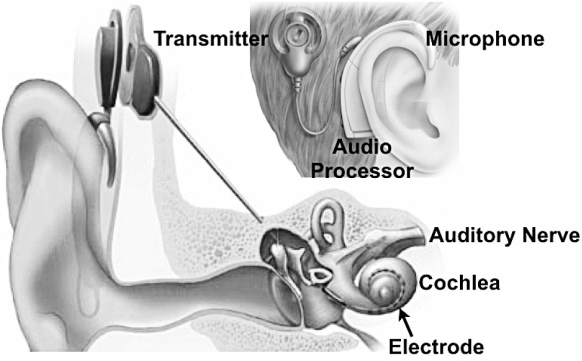


Diagram of a cochlear implant. The sound captured by a microphone is processed and transmitted to an electrode that electrically stimulates the cochlea, creating nerve impulses that are, in turn, transmitted to the brain through the auditory nerve.

Several clinical trials have shown how these retinal implants can allow patients with blindness due to retinal degenerative diseases to detect movements and simple shapes—such as a square or triangle—and move around in a room without crashing against objects.[17] These advances are promising, but we still have to get to the point where patients could read or recognize faces and more sophisticated objects in order for retinal implants to be used on a massive scale. The big problem is that there's an incredibly vast repertoire of things we recognize with our sight, which trigger different patterns of neural discharges in the optic nerve. In other words, sufficiently specific and distinguishable stimulation patterns need to be created in the retina to be able to make out the objects in front of us. Clearly, the number of patterns that can be identified will depend on the number of microelectrodes implanted (the more electrodes, the greater the resolution and ability to discriminate). In that sense, the remarkable advances in the development of miniaturized electrodes during the last few years allow us to foresee great progress in the use of retinal implants in the near future.[18]

Like sound, color doesn't exist around us; what does exist are electromagnetic waves, rays of light that impact our retinas. The color and the shape of everything we see are constructions in the brain based on the interpretation of nerve impulses created in the retina. So, if sight, like hearing and other senses, is nothing more than the interpretation of nerve impulses, nothing stops us from capturing the electromagnetic waves and transmitting them to the brain in some other way.

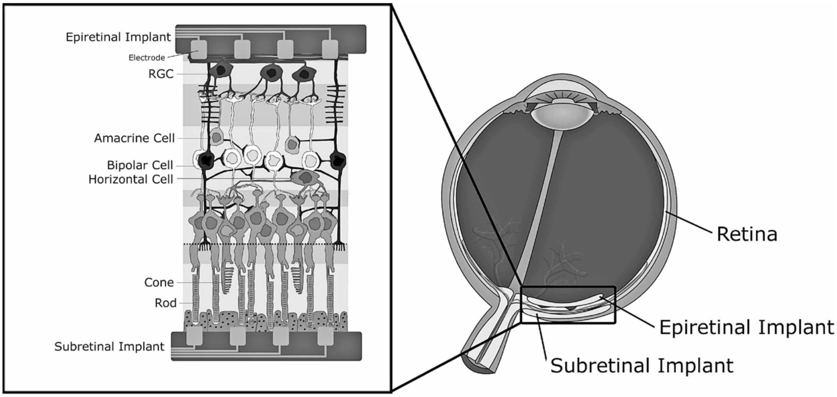


Diagram of a retinal implant. The stimulation through electrodes can be done in the layer that holds the retinal photoreceptors (subretinal implant) or directly in the layer of RGC (retinal ganglion cells) neurons that make up the optic nerve (epiretinal implant).

In the late 1960s, Paul Bach-y-Rita developed a system by which images that were recorded with a camera generated a stimulation pattern on the backs of blind people.[19] With this system, participants were able to recognize different everyday objects (a telephone, a chair, a cup, and so on) and capital letters, among other things. However, the back does not possess a high density of receptors, and follow-up developments gave rise to the BrainPort system, which uses stimulation patterns on the tongue, from the signals recorded by a camera mounted on special eyeglasses. (The advantage of using the tongue is that it is one of the organs with the most nerve receptors in the human body.) And just as BrainPort converted visual information into tactile stimuli, systems such as vOICe convert visual stimuli into sounds that allow someone to differentiate objects. In sum, by recording the electromagnetic waves that make up the visual stimuli, different methods can be implemented to transmit information to the brain and allow a person to see—provided, of course, that the brain can interpret such information.

Analyzing the issue brought up by his writer friend William Molyneux, John Locke wondered about the perception a person who was born blind would have when seeing for the first time (for example, when observing a sphere or cube). Locke concluded:

> [I] am of opinion that the blind man, at first sight, would not be able with certainty to say which was the globe, which the cube, whilst he only saw them; though he could unerringly name them by his touch, and certainly distinguish them by the difference of their figures felt.
>
> —John Locke, *Essay Concerning Human Understanding* (Book II: ch. 9, sec. 8), 1690

Similar arguments, denying the possibility of innate knowledge and highlighting the importance of experience, were outlined by Bishop George Berkeley in his *Essay Towards a New Theory of Vision*, published in 1709, and by Hermann von Helmholtz, who claimed that "memories of previous experiences have a preponderant and influential role in our visual observations."[20] So, it is not surprising that people who start using the BrainPort or vOICe systems, or those who receive retinal implants, need plenty of training to begin to recognize objects. Merely connecting a sensory organ to the brain is not enough because the brain needs to learn how to interpret these signals. However, the neural plasticity we previously spoke about makes this eventually possible.

We can not only substitute, but also increase our sensory abilities. I'm not talking about having infrared vision like serpents (or the alien in the *Predator* movies) or hearing high frequencies like dogs; I'm talking about the possibility of developing new senses. This is not unheard of in the animal kingdom. Some fish, especially in the Amazon, can detect electric fields (and sense alterations in the field they themselves emit to locate prey or predators); some birds also detect magnetic fields, which help them get a feeling of where North is, to fly in the right direction, and so on. It follows that we could then ask ourselves what type of information we would like to *feel* directly in our bodies. And just as we see different colors, we could detect the intensity of Wi-Fi bands—for example, through the stimulation of tiny electrodes in our forearm—to find the spots with the best internet connection. With similar sensors, we could also feel variations of stock markets in which we have invested our money; the number of mentions on Twitter that involve us; the number of likes on our Facebook posts; the number of online sales of our new book while we're promoting it; or the rating of a radio or TV program in which we are involved.* It wouldn't be strange if this direct perception of our financial situation or our popularity in the media were to end up creating different cerebral reactions—for example, through discharges of adrenaline or activations of reward areas—as they naturally are generated through signals coming from the senses.

The possibilities for increasing our senses are incredible. Although I'm far from suggesting something like that. The internet and gadgets of our time are already overwhelming us with information day in and day out, and amplifying our senses would only exacerbate this problem. It would be like being permanently online, as is the case with many teenagers who wander through their lives hypnotized by their cell phone screens; it would mean removing the focus from what is most important: exercising our unique ability to think, to make abstractions and extract a meaning from things, to associate these things with others and assign them context. In fact, it has been estimated that nowadays we are exposed to an amount of information equivalent to that contained in 174 newspapers a day, five times more than in the 1980s.[21] Such an onslaught of information goes against what makes us human and distinguishes us from other animals or computers. This is an incredibly interesting and topical subject, but right now I'll end this discussion about the information overload we are exposed to because I've already ranted enough about it—particularly against the use of smartphones and being permanently online—in two previous books.**

So, we've seen how advances in recent years allowed us to catch a glimpse of the very real and near possibility of using neural prostheses to restore the function of different limbs and replace or even increase our senses. Through extraordinary developments, what was considered science fiction only a few years ago is already being tested with humans. The discussion on replacing different parts of the body led us to wonder about the problem of identity, defining that a person is who they are and not someone else. We also saw that identity can be defined through the body or through memory, although this distinction actually doesn't make much sense since it implicitly assumes a dualist view. After all, the brain is also part of the body and memories are nothing more than the activity of interconnected neurons. However, we still have to discuss the possibility of a much more delicate and controversial organ replacement.

If we received the heart of a person, let's call him John, we would continue to be the same person but with a new heart; if we received his liver, we would still continue to be the same but with another liver; the same with the pancreas, intestines, eyes, and we could go on with every organ, with a great exception: If we received John's brain (something that for now is impossible, but let's just imagine for a moment that it was possible), then we would no longer say that we continue to be the same person with a new organ; we would instead say that John received a new body. As we previously argued, the brain is the warehouse of memory and identity. And if instead of replacing our brain we could preserve it, or replicate it, could we perpetuate ourselves beyond death? This is the big question with which we will close this book, but on the way there, let's next talk about memory and the influence that our dreams have on it.

---

* This text may bring back some good old memories to those who grew up in the 1970s. *The Six Million Dollar Man* ran for five seasons, from 1974 to 1978. Numerous compilation clips and episodes can be found online.

* Not to be confused with Richard Ander*son*, the actor who played Goldman in *The Six Million Dollar Man*.

* Richard's discovery happened while performing experiments with monkeys. To review these results, see Richard A. Andersen and Christopher A. Buneo, "Intentional Maps in Posterior Parietal Cortex," *Annual Review of Neuroscience* 25, no. 1 (2002): 189–220.

* It is no coincidence I'm using the late Christopher Reeve as an example. After his accident and with extraordinary willpower, he created a foundation to support the development of neural prosthetics. Many of the studies in this field have been at least partially funded by his foundation. As proof of his tenacity and hope, up until his death in 2004, he was convinced that thanks to the advances in science he would walk again. Sadly, he would have had to live a few more years to see his dreams come true.

* By far, patients with paralysis care less about walking again than about being able to use their hands to reach objects, feed themselves, and carry out another slew of tasks for which they constantly need help. See, for example, Kim D. Anderson, "Targeting Recovery: Priorities of the Spinal Cord–Injured Population," *Journal of Neurotrauma* 21, no. 10 (2004): 1371–1383.

* This is due to several factors, among them, small movements of the electrodes and changes in impedances generated by blood healing around them.

* Actually, not all neurons modify their activity. To optimize results, neuroplasticity is complemented with the use of adaptive algorithms, where parameters are adjusted to improve performance. For more information, see Amy L. Orsborn et al., "Closed-Loop Decoder Adaptation Shapes Neural Plasticity for Skillful Neuroprosthetic Control," *Neuron* 82, no. 6 (2014): 1380–1393; and Krishna V. Shenoy and Jose M. Carmena, "Combining Decoder Design and Neural Adaptation in Brain-Machine Interfaces," *Neuron* 84, no. 4 (2014): 665–680.

** In fact, the results for the patient pictured on page 159 were obtained five years after implanting the electrodes in her brain. For more information, see Leigh R.Hochberg et al., "Reach and Grasp by People with Tetraplegia Using a Neurally Controlled Robotic Arm," *Nature* 485, no. 7398 (2012): 372–375.

*** The *RoboCop* remake is filled with philosophical nods. Dr. Norton's first name is a clear reference to Daniel Dennett, whom we discussed in earlier chapters. In the movie, Senator "Hubert Dreyfus" opposes the use of robots for security control, in line with the view of a philosopher by the same name, who has been a prominent critic of the use of artificial intelligence.

* The myth follows with a tragic twist: In the joy of his triumphant return, Theseus forgets to hoist a white flag, as he had promised his father, King Aegeus, in case he successfully came back. Aegeus climbed a promontory every day to search the horizon, waiting for news from his son. When he saw the vessel returning without the white flag, he assumed Theseus had been killed by the Minotaur and threw himself into the sea, hence the name Aegean Sea.

* This is what Plato writes at the beginning of *Phaedo* to explain the delay of Socrates's sentence (see chapter one), since executions were prohibited during the journey of Theseus's ship to the Island of Delos.

** One of the most famous sorites paradoxes is the one of the bald man. If we add one hair at a time to a bald man's head, after each hair added, we will continue to say he is bald, regardless of how much hair he will eventually end up with after repeating this procedure over and over again.

The discussion about the perception of change is deeply rooted in Greece's ancient philosophy, starting with the contrary positions of Heraclitus, for whom everything changes, and Parmenides, who claimed that the perception of change is unreal. Aristotle later reconciled both positions in his theory of actuality and potentiality (Aristotle, *Metaphysics*, Book IX).

* A similar reasoning was proposed by the British philosopher Bernard Williams, who considered the case of the crazy scientist who exchanges our mind with that of someone else's. The scientist tells us he is going to torture one of the two people, and he lets us choose: Do we prefer he torture the other person's body, but with our mind, or our body with the other person's mind? If we choose the first option, we favor the idea that identity is tied to the body, while if we choose the second one, we believe identity goes with the mind, with memory. Bernard Williams, "The Self and the Future," *The Philosophical Review* 79, no. 2 (1970): 161–180.

* It should be noted that this statement is also tricky and isn't necessarily accepted by all philosophers. In fact, the discussion about what determines identity has been going on for centuries in philosophy. For a more detailed description, see, for example, chapter six in John Hospers, *An Introduction to Philosophical Analysis* (London: Routledge, 1956).

* Here is one of many YouTube videos that shows the hilarious reactions of subjects when, after making them feel that the rubber hand is part of their body, it is hammered or pricked with pins: "The Rubber Hand Experiment" YouTube video, 2:15. Posted by Spidy, October 31, 2017. www.youtube.com/watch?v=RaP0MqvkvUw.

* I met Miguel at a conference in Los Angeles. After a few minutes of a somewhat formal exchange, his demeanor noticeably changed when we began to talk about soccer. To my surprise, Miguel knew the second league coaches in Argentina and made me laugh out loud with endless anecdotes about the Copa Roca that Argentina and Brazil used to play. The least enjoyable memory of that encounter was, however, the fact that he usually began his talks describing the complexity of visual motor coordination with one of Brazil's free-kick goals against Argentina.

* The results showing a paralyzed patient moving his legs and walking with a walker were obtained by Grégoire Courtine's team. Regardless of which group becomes the first to make a paralyzed patient walk with a neural prosthesis—something that is close to becoming a reality—I see this as an advance of science in general, supported by a great number of results by several groups, mainly in the last five to ten years. Up until now, stimulation in humans to restore locomotion has been open loop, using continuous stimulation to recover part of a member's mobility. See Claudia A. Angeli et al., "Altering Spinal Cord Excitability Enables Voluntary Movements After Chronic Complete Paralysis in Humans," *Brain* 137, no. 5 (2014): 1394–1409.

What still needs to be done is to close the circuit and make the stimulation start with the activity of the neurons in the motor cortex, as already done in monkeys.

* I mention "sound as we know it," referring to the sound produced by variations in air pressure. Of course, sounds can also be transmitted through solids, liquids, and gases.

** A cochlear implant is very different from an earphone, which does nothing more than amplify the sound recorded by a microphone. A cochlear implant, on the other hand, transforms sound into electric impulses. For more information, see J. P. Rauschecker and R. V. Shannon, "Sending Sound to the Brain," *Science* 295, no. 5557 (2002): 1025–1029.

* Some of these ideas were presented in a very entertaining TED Talk by scientist/author David Eagleman in March 2015.

** *Borges and Memory* and *The Forgetting Machine* (cited in chapters one and two, respectively).

CHAPTER 8

INCEPTION

Dream Construction

I*nception* (2010) is arguably Christopher Nolan's masterpiece, and that is no small thing coming from the director of *Memento, The Prestige, Dunkirk*, and *The Dark Knight* (Batman) trilogy.

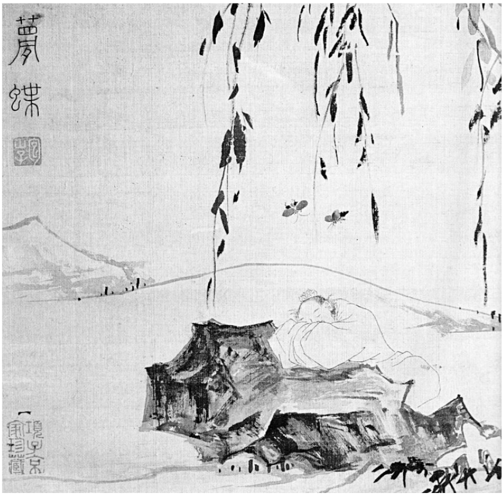
A group of specialists led by an "extractor" named Cobb engage in corporate espionage, stealing secrets from people's subconsciouses while they are sleeping. The idea is that while awake, we have defense mechanisms to protect our secrets, but during sleep, these mechanisms diminish and our thoughts are susceptible to theft. Cobb and his henchmen create "shared dreams," where they can enter and prompt the victim to reveal the secret they want. However, to avoid discovery, the extraction must be subtle and masked within different levels of dreams. And so the plot immerses us in the fascinating world of dreams, dreams within dreams, and how these articulate and interact with each other: For example, a falling car in one dream level creates an instant lack of gravity in the dream inside the dream, and a snow avalanche in another even deeper dream level.

The movie integrates five realities—each one in successive dream levels, with their own space and time—and is chock-full of reflections on what we experience while we sleep. Dreams violate the laws of physics and expand time, which means we could spend hours falling without ever hitting the ground. Further, the dream world is unstable and lacks narrative coherence; both dreams and wakefulness are constructed by the brain, but in dreams this construction takes place exclusively through imagination, while when awake, our reality is constructed from sensing our surroundings.

In several interviews, Nolan admits being a Borges fan and highlights the author's influence on the plot. He specifically mentions "The Secret Miracle,"[1] where a writer of Jewish descent pleads with God to give him one more year of life to finish his masterpiece, before being shot by the Nazis after the invasion of Prague. God agrees, but in a very peculiar way: he stops time for a year right before the writer is shot and there, in front of the firing squad aiming straight at him, the man reviews and mentally corrects his work's plot until he finally finds the words for the last epithet, and at that instant, time resumes and he is shot. A similar dilemma is raised in *Inception*, where one of the conspirators is fatally wounded, and Cobb's team decides to take him into a deeper dream level to prolong his life, since time slows down there.[*]

Another clear Borges influence is "The Circular Ruins," first published in 1940. Like the "architects" who design dreams in Nolan's movie, Borges's story describes a character who constructs a person in his dreams. The monumental task, imagining each organ, each feature, and even each individual hair strand, takes years. After achieving his goal, he welcomes the approach of a titanic fire, believing that death will release him from his labors. But upon noticing that the flames aren't burning him, he realizes that he himself is being dreamed up by someone else.

Borges repeatedly plays with the apparent difficulty of distinguishing between reality and dreams—a quandary that has been discussed in philosophy for centuries. In fact, one of his anthologies, *Libro de sueños* (*The Book of Dreams*), includes a famous story by the ancient Chinese philosopher Zhuang Zhou, who was already raising this issue more than two thousand years ago: “One day I dreamed I was a butterfly, and when I woke I didn’t know if Zhuang Zhou had dreamed he was a butterfly, or if a butterfly was now dreaming it was Zhuang Zhou.”



Zhuang Zhou’s butterfly dream, by Chinese artist Lu Zhi (sixteenth century).

Like Borges and Zhuang Zhou, *Inception* makes us question reality and ask ourselves how we know we aren’t dreaming. Nolan says he was inspired by movies like *The Matrix*, where humans live in a fictitious reality created by a powerful artificial intelligence. *The Matrix* challenges the existence of reality, while Nolan’s *Inception* reverses the question and wonders if we could create an alternative reality in someone else’s mind (where we could immerse ourselves to steal their secrets). Yet when discussing the possibility of being in a dream, first we must differentiate two alternatives: we can ask ourselves if *our entire life* is a dream, or we could take a less dramatic stance, accept that reality exists, and ask ourselves if *at this instant* we are dreaming.

The first version, doubting reality, is what Descartes brought up in his *Meditations* (the first one), asking himself if a devious and powerful evil genius could make “the sky, air, earth, colors, figures, sounds, and all other external things be just illusions and dreams” to deceive us. Closer in time, Descartes’s idea is the *brain in a vat* argument we discussed in chapter four, when we talked about *The Matrix*. However, from Dennett’s arguments, we saw that a fictitious construction of reality cannot include the infinite details of our interactions with the world, and sooner or later the deception would be obvious. (In fact, one of the people Cobb aims to dupe in *Inception* realizes he is in a fictitious reality when he notices that the rug in his office is not made of wool, like the original one, but polyester.) We also saw that Ludwig Wittgenstein showed that if the external reality didn’t exist, if we didn’t have other people to interact with, knowledge would be impossible.

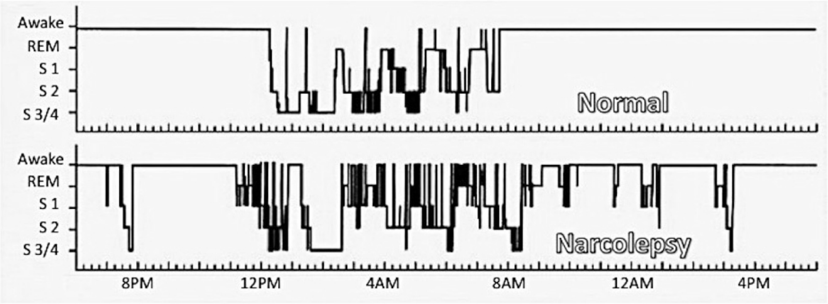
Accepting, then, that reality exists and that our life cannot be just a dream, let’s analyze the other alternative: What evidence do we have that we are not dreaming *right now*? In other words, how can we differentiate dreams from wakefulness? Nolan offers an ingenious solution: Cobb always carries with him a unique totem—the spinning top featured in this chapter's opening illustration—that only he knows in detail; so, when faced with the question of whether or not he is in someone else's dream, all he has to do is check his totem to see if it is the same as always or a re-created copy inside a dream.* (Of course, this only makes sense within the movie's context, since we aren't usually carrying around totems that we could hold on to in our dreams.)

For most people, distinguishing between dreaming and being awake seems trivial, almost obvious, but in some cases it is not so easy. When we sleep we go through different stages, which can be classified by stereotyped patterns in electroencephalographic recordings. A normal person has cyclical sleep stages throughout one night (alternating approximately every ninety minutes), while people with sleeping disorders have a much more disorganized transition between each stage. This is the case of patients with *narcolepsy*, a sleep disorder characterized by excessive sleepiness during the day, where brief sleep stages appear amid wakefulness. The lack of a clear distinction between sleep and wakefulness makes patients with narcolepsy believe that experiences in dreams actually happened in reality, sometimes causing dramatic situations. Examples abound: A student vividly recalled stepping off a bus and being picked up by a person who abused her in a car; however, after several testimonies and psychological analyses, police were able to determine that she had fallen asleep on the bus and had dreamed of this event. A married woman was disconsolate because she believed she'd had an extramarital affair, which she later ascertained was only a dream. Likewise, a man reported a neighbor for constantly throwing garbage in the trash can at his home and later had issues with his boss after believing he'd had a sexual relationship with the other man's wife.[2]

Aside from narcolepsy cases and some other specific pathologies (such as schizophrenia), it is highly unusual for someone to confuse sleep with wakefulness.[3] But what allows us to naturally make this distinction?

In his *Meditations on First Philosophy*, Descartes argues that "there are not enough concluding indications and signs to clearly distinguish sleep from wakefulness." However, he later concludes that the continuity of experiences is what distinguishes wakefulness from sleep.

> Well, now I see a noticeable difference between them [wakefulness and sleep]: and it is that our memory can never link and join our dreams together, nor the course of our life, as it is in the habit of joining things that happen to us while awake . . .
>
> . . . when I perceive things, knowing distinctly where they are coming from and where they are, as well as the moment they appear, and being able to uninterruptedly weave the feeling I have for them with the rest of the course of my life, then I am sure that I am perceiving them awake and not asleep.
>
> —René Descartes, Sixth Meditation[4]



Stages of sleep and wakefulness in a normal subject (*above*) and in a person with narcolepsy (*below*).* The person with narcolepsy has much more disorganized transitions between different sleep stages (REM sleep, and stages 1, 2, 3, and 4) than the normal subject, in addition to sleep stages that also occur during wakefulness.

Thomas Hobbes, Descartes's contemporary, was very critical of this last argument. Hobbes reasoned that in principle one could also dream that what is happening in a given moment is related to past events and, therefore, the principle of continuity wouldn't distinguish sleep from wakefulness.[5] However, beyond Hobbes's objection being theoretically possible, sleep lacks temporal continuity. In my dreams, I can be in my office writing this book and then on an elephant traveling across India. The lack of external stimuli causes me to lose coherence with reality, yet while awake, time progresses linearly, and I can't go from one place to another without having traveled through the intermediate points; what happens to me in a particular moment is intrinsically tied to what just happened to me, and what will happen in the following minutes.

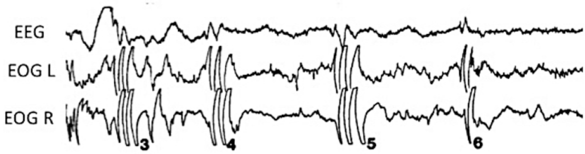
We are sure we do not dream while awake, but as Nolan argues in *Inception*, our dreams also appear real while they last, and we only notice their incoherence when we wake up. The issue is that when we sleep, our state of consciousness is very limited, and we aren't able to ask ourselves if what we are experiencing is a dream. However, that's not always the case.

In what are known as *lucid dreams*, we can ask ourselves these questions and eventually realize we are not awake. *Inception*'s plot is precisely based on secret-stealing thieves who are conscious that they are in a dream where they carry out their deceptions.

While sharing a dream for the first time with a new member he has just recruited to his team, Cobb says that we never remember the beginning of dreams; we simply appear in a situation without knowing how. Right then, sipping a coffee in Paris, he asks the student, "How did we end up at this café?" The young woman seems confused, doubts for a moment, and finally realizes she's dreaming. To test reality, there are certain types of questions we can ask ourselves while we sleep. For example, we can wonder how we reached a certain place; we can challenge the laws of physics and try to penetrate objects with our hands, or remain suspended in a midair jump; we can cover our nose and try to breathe through our closed mouth; we can read some lines of text to see if they make sense.* With these tests, we could verify if we are having a lucid dream.

Roughly one in two people state they have had at least one lucid dream in their life, and approximately one third of these have more than one lucid dream a month.[6] But the idea of lucid dreams is not that easy to accept; after all, sleeping is usually described as a state where we lose consciousness. Therefore, besides subjective reports, what scientific proof do we have to confirm that a person is conscious of being in a dream? Psychologist Stephen LaBerge set out to find such proof during his PhD work at Stanford University, something seemingly impossible since common sense tells us that, in principle, we cannot communicate with somebody during a dream so they can report being conscious of what is happening.

During sleep we make eye movements, which produce variations in the electric field that can be captured by an electrooculogram (EOG).* Using this principle, LeBerge asked subjects who said they had lucid dreams to do preestablished ocular movements (from left to right) as a way to testify if they were conscious during their dream. Once he observed these specific eye movements in the EOG, he just had to wake the subject and corroborate that the recorded EOG patterns corresponded with the ocular movements the person intentionally made while dreaming.



Electroencephalogram (*above*) and electrooculographic (left eye movement, *middle*; right eye movement, *below*) recordings in a subject who made several preestablished movements (marked as 3, 4, 5, 6), reporting that he was lucid while dreaming.**

Going back to *Inception*'s premise, the key to extracting secrets from victims is based on the possibility of sharing dreams and hence accessing the subconscious. We saw in chapter five that up to a certain point, reading a person's dreams is already a reality: three years after *Inception* was released, Yukiyasu Kamitani's team at Kyoto University showed that it was possible to predict what someone was dreaming based on the recordings of cerebral activity measured with fMRIs. But could we share a dream?

To get two people to eventually share a dream, we need their brains to interact and communicate. In other words, one person's thoughts must flow directly into the other one's through what is called *brain-to-brain communication*. An experiment of this sort was carried out at the lab of Miguel Nicolelis (whom we introduced in the previous chapter). A rat was trained to recognize stimuli that indicated what lever to push to obtain a reward (a drop of juice), while its cerebral activity was recorded and transmitted to a second rat (stimulating its cerebral cortex) located in another cage, which could not see the stimuli but could still correctly decide what lever to push according to the cerebral activity it received from the first rat.[7] However, it is important to mention that there's a huge difference between transmitting a binary decision (choosing a lever) and transmitting a dream's content. For the latter to be possible, we would have to selectively stimulate specific groups of neurons in a subject based on the activity recorded in another one. For example, if one subject dreamt of a fuchsia sky, we would have to interpret the content of this dream and then identify and stimulate the neurons that represent the sky and the color fuchsia in the other subject. This is already extremely complex, and we would only be transmitting the color of the sky, forget about the specific contents of a dream.

Setting aside the possibility of manipulating another person's dreams, which to this day is still sci-fi, we now raise a much more tangible but no less spectacular question: Could we manipulate or even construct our own dreams?

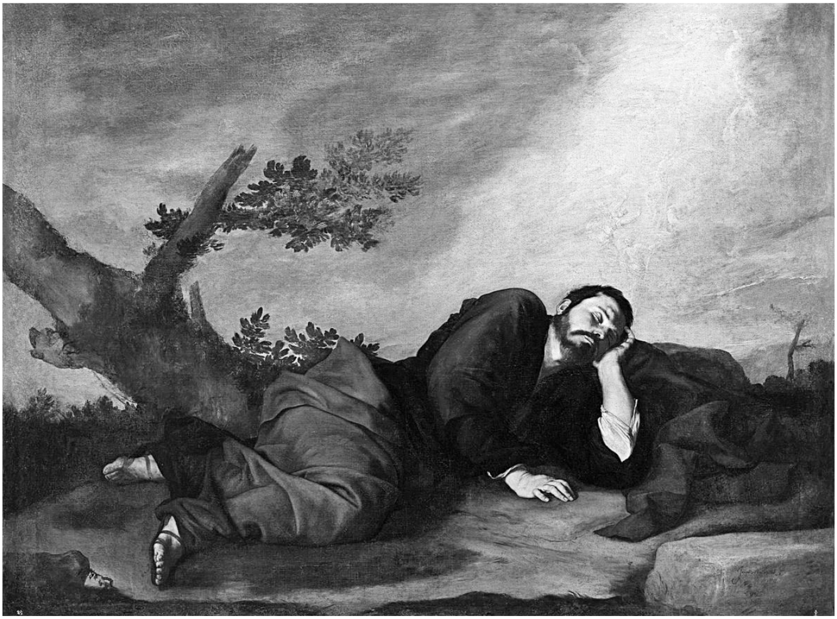
We have referred to lucid dreams. There are countless courses, books, and web pages dedicated to this subject. The reason is simple: If we are conscious that we are dreaming, we could, for example, avoid a nightmare, since knowing something is not real should dissipate the eventual fear it would create in us. But even more interesting is the fact that, while aware of our dreams, we could be capable of making decisions to manipulate them. For example, we could choose not to enter a scary-looking house where we know endless problems await us and decide, instead, to take a great leap and magically float in the air—as we usually do in our dreams. If we could conquer our dreams, the possibilities would be endless: We could travel, get together with people we want to meet, fly, go to other planets, reach the bottom of the ocean. The possibility of influencing what happens in a dream is without a doubt incredibly attractive. But is it truly possible to construct a tailor-made dream like in *Inception*? Could we choose what to dream before going to sleep and meet up with our favorite actress or a loved one, or travel to the moon?

The attempt to influence our dreams is known as *dream incubation*. We usually tend to dream about things that happen to us during the day. Dream incubation consists of thinking over and over about what we want to dream before falling asleep. Deirdre Barrett, a psychologist at Harvard Medical School, tested this hypothesis by asking a group of students to try to "incubate" dreams during the course of one week, using any topic they considered relevant. The students then wrote down what they remembered of their dreams and approximately half of these recollections were related to the chosen subject.[8] Barrett's results aren't too surprising considering the experience we generally have of continuing to think about the problems that worry us while we're awake (sometimes finding creative and innovative solutions). However, it is one thing to influence the content of dreams and another very different one to control what is going to happen in them.

Dreams have their own dynamic, which we cannot voluntarily choose at will. Regardless of how easy it could be to have lucid dreams or even to incubate dreams, we could possibly manipulate the general theme of a dream, but not its specific plot. Although, of course, determining what we are going to dream about is already a great feat—I, for one, would rather dream about something pleasant than end up in a nightmare.

We spend a third of our lives sleeping, no less than twenty-five years for a seventy-five-year-old person. We dream we are falling, that we are being chased, that we don't pass an exam, that we are naked in public, that we can't run, or that our teeth are falling out. But why?

In ancient times, dreams were believed to be premonitions. The Old Testament, and particularly the Book of Genesis, describes a large number of dreams in which God communicates with men. The painting shown on the next page represents Jacob's dream (from Genesis 28:12): Having deceived his brother and fearing his vengeance, Jacob decides to leave home. Along the way, spending a night in an open field, he dreams of a ladder reaching up to heaven on that very spot, by which angels come and go. From on high, God manifests that on that land he will protect Jacob and all of his descendants, and, according to the Hebrew scriptures, when Jacob awakes, he decides to make this land his home and names it Bethel.

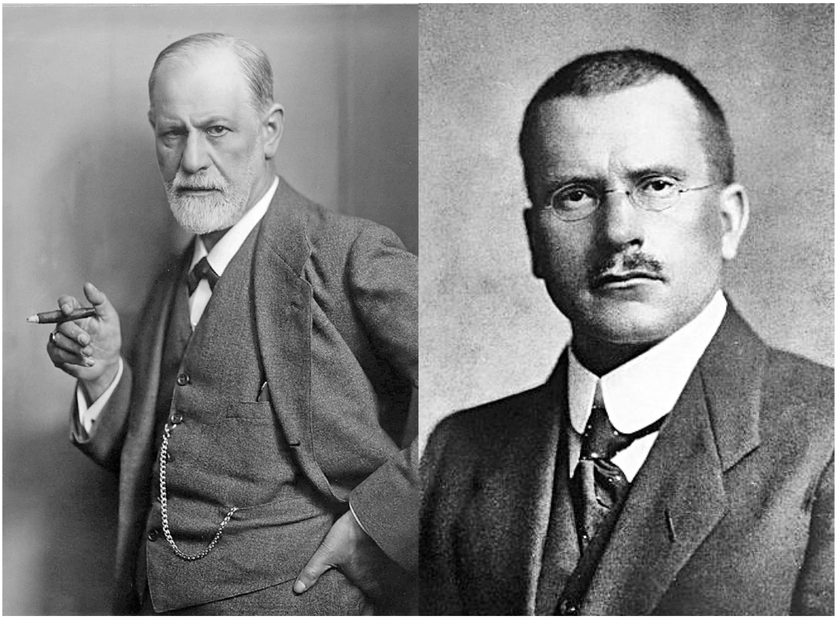


Jacob's Dream. Oil painting by José de Ribera (1639).

Beyond these stories, there is no scientific evidence that supports a premonitory role of dreams and, even if we are prone to believe the few testimonies of dreams anticipating the future, they are so rare that it is reasonable to attribute them to random coincidences. It is much more logical to argue that sleeping is the rest that our bodies, and particularly our brains, need after a long day. But this doesn't justify dreams either. In fact, we would probably rest much better if our minds were completely blank while we slept, and there would be no reason to create the extraordinary worlds that we barely remember when we wake up.

The rigorous and systematic study of dreams began with Sigmund Freud (1856–1939), who would lay the foundations of psychoanalysis. In his most famous work, published at the start of the twentieth century, he says the "interpretation of dreams is the real path toward knowledge about the unconscious activities of the mind."[9] Freud argued that dreams do not have external causes (the gods handing us different prophecies) but internal ones (our subconscious), hence his interest in studying them. Freud saw dreams as an escape valve that unleashes the desires (many times of a sexual nature) that we repress while awake, because when we are not sleeping, we have censorship mechanisms that inhibit us from going ahead with embarrassing actions.[10]

Therefore, dreams unleash forbidden desires in our subconscious (those we forget when we wake up), relieving the frustration of not being able to act on them while awake.



Two pioneers in the study of dreams: Sigmund Freud (*left*) and Carl Jung (*right*).

Currently, well over a century after he wrote his book, it is easy to criticize Freud. To put this into context, when Freud published *The Interpretation of Dreams*, Spanish neuroscientist and pathologist Santiago Ramón y Cajal was just unveiling his doctrine considering the neuron as the basic unit in the brain, while nowadays the advances in neuroscience allow us to measure the activity of thousands of neurons and see how they determine different behaviors.

Freud held arguments that we now know not to be true, but the important fact is that, right or wrong, he created a discussion that gave way to remarkable advances in our knowledge. However, this is no excuse for certain branches of psychology, and particularly psychoanalysis, to continue worshiping dream interpretation "à la Freud."[11]

Setting aside possible errors in the implementation of Freud's ideas in psychoanalysis, the criticism of his theory of dreams is much more concrete. Freud posited that dreams have a liberating effect because if we didn't dream—hence if we were unable to satisfy our repressed desires—we would go mad, something that later was proven not to be the case.

In the early 1950s, Eugene Aserinsky was a physiology student at University of Chicago, carrying out his PhD with Nathaniel Kleitman, perhaps the most renowned dream expert of his time. The topic of Aserinsky's thesis was to record, during endless hours, EOGs in subjects to study the movement of their eyes while sleeping. (Remember that while we are sleeping, every so often, our eyes move involuntarily.)

Until then, it was believed that sleep was no more than a uniform period of rest, for the brain to recover energy and confront a new day. However, doing EOG recordings with his eight-year-old son, Armond, one day Aserinsky observed periods of sleep where the eyes paradoxically moved like when we are awake, hence the name *paradoxical sleep* or REM (rapid eye movement) of this sleep stage.[12] The unexpected discovery would radically change the idea we have about dreams, which we now classify as REM and NREM (non-rapid eye movement).* Ocular movements during the REM stage can be easily seen with the naked eye (even though the sleeper's eyelids are closed) and, surprisingly, during thousands of years, the greatest thinkers were interested in dreams, but never before had anyone noticed those eye movements.

The next question Aserinsky asked himself is obvious. What happens during REM sleep? Why do we have those periods when we frantically move our eyes back and forth? There's a more than tempting possibility: If dreams happen specifically during REM sleep, ocular movements could be due to eye fixations as we explore the whereabouts of what we are living in our dreams. Aserinsky decided to wake subjects while they were sleeping, and this way he proved that the vast majority of dreams were indeed happening when he woke them from the REM stage. Nowadays, however, we know that ocular movements do not correspond to visual explorations in our dreams. Moreover, we know that dreams also happen in the NREM stage, although with a much smaller probability and without the emotional weight and bizarre narrative constructions that occur during REM sleep.[13]

We have taken a long detour talking about the ins and outs of the discovery of REM sleep and its relationship with dreams because with that information we now have all the tools to experimentally test Freud's theory. If dreams are a person's liberation of repressed desires, depriving them of their dreams—waking them each time they enter the REM stage—would have devastating effects. However, although the deprivation of REM sleep does cause certain minor disorders, it certainly does not have the consequences predicted from Freud's theory.[14]

More consistent with these results is the view of Carl Jung (1875–1961), Freud's disciple and (at first) friend, who took a different position from his mentor on dreams (which eventually destroyed their friendship). For Jung, instead of hiding repressed desires, dreams are a window into our unconscious and have a counterbalancing effect, making unconscious perspectives complement conscious wakefulness experiences. Moreover, Jung also argued that dreams provide visions of possible futures—in other words, not just premonitions, as believed in ancient times, but constructions that prepare us to confront scenarios that could happen. Meaning, if at some point we face a new situation in our lives, it is possible that we unconsciously know how to react because we have experienced similar circumstances in our dreams.

A radically different view on dreams was proposed in the late 1970s by Allan Hobson and Robert McCarley, two psychiatrists from Harvard University.[15] For these researchers, during sleep, the activation of groups of neurons, and what they represent, is generated randomly. For example, let's imagine that neurons activated by chance represent the concepts of sea, tree, buffalo, and castle—activations that are likely related to things experienced during daytime. But what does this have to do with dreams? The idea is that, through these activations, the brain builds a narrative. For example: "I'm at the beach and there's a big tree in the sand and a buffalo runs after me, but never reaches me, until I finally lose it when I get to a castle." The story makes no sense, just like dreams don't make sense either. Hobson and McCarley believed that dreams are so strange because the brain generates a story based on random neural activations.

When speaking about Helmholtz and Bartlett, we saw in chapter four that both perception and memory are brain constructs. The difference with a dream is that while awake, these constructions are formed from coherent stimuli that are related to each other because they respond to reality (in real life, I will not see a huge tree in the middle of the beach, a buffalo, and a castle), but our perception of what is happening both in a dream and while awake is based on the brain’s unique ability to construct a reality from very little information. We do this unconsciously all the time; it is so automatic and inevitable that we see, for example, constellations representing mythological heroes in stars randomly distributed in the firmament; it is so misleading that Las Vegas casinos show the last few dozen roulette results, so naïve players believe they have found a pattern that in reality doesn’t exist and proceed to lose their money.

A German proverb says *Träume sind Schäume* (dreams are but froth). In other words, dreams are nothing, like soap bubbles, or as more formally expressed: they’re an *epiphenomenon*. An epiphenomenon is, for example, a car engine’s noise. The function of the engine is to move the car, not to make noise; the noise is a secondary effect. Likewise, according to Hobson and McCarley’s theory, dreams have no function, they’re just the harmless consequence of random activations in the brain.

It looks like neuroscience finally gave Freud and Jung’s dream interpretations the coup de grâce. However, I think this assessment is somewhat unfair. Freud and Jung have the merit of being pioneers in seeing dreams as a tool to access the unconscious. Although dreams don’t hide repressed sexual desires and are not particularly useful in compensating for our conscious experiences with unconscious fantasies, they do give an account—as manifested in the dreams themselves—of our emotions and interests, which are the driving forces underlying our impulses. These impulses, in turn, underlie the feeling of free will that drives our decisions, an apparent exercise of our volition given by subconscious drives.

The view of dreams as epiphenomena is also problematic because, once again, it implicitly assumes Cartesian dualism, the division between mind and brain. By saying that dreams are epiphenomena, we are saying that what does exist are the neural activations, whereas what they represent—and the construction we make from these representations, meaning, a dream—is nothing more than a side effect. However, as stated by the materialistic view we have been supporting from the start, neural activations and what they represent are one and the same thing. Materialism brings us once again to the hard problem we described in chapter two, which is worth briefly revisiting.

Let’s imagine, one more time, that while we sleep we activate neurons that represent the sea. There is no salt water in our brain, only small electric impulses that are transmitted from one neuron to another. These electric impulses are a representation of reality, not reality itself, just as the sequence of zeros and ones etched onto a CD represent different pieces of music. Clearly, this binary code is illegible to the naked eye, but we also use other representations, like the notes on a music sheet.

ff
p
cresc.
10
cresc.
f
cresc.

The first few bars of Beethoven's Fifth Symphony. The notes represent a piece of music, as the activation of groups of neurons represent the external reality in our brain.

A musical score rendered on a piece of paper—like a binary sequence of numbers—is physically not music, but it represents it. However, we could say there's a fundamental difference between the music score and the activation of neurons in the brain. The score is actually an "instruction manual" to manipulate instruments that will eventually execute music. The relationship between the music we hear and the score—or the sequence of zeros and ones that we decode via a CD player—is very clear. The "hard problem" in neuroscience is understanding how, through the activation of neurons, we *feel*, for example, the sea, its waves, the smell of seaweed, and the intense cold when we dive in (at least in the waters of the Argentine Atlantic coast). The relationship between neural activations and what we feel isn't as obvious as in the case of the piece of music. But these feelings—again, what we call qualia—are merely activations of groups of neurons that represent things that are connected to each other in our brain: When we think of the sea, we activate a group of neurons that represent this concept, which in turn activates other groups of neurons representing the smell of seaweed, the sight of waves, the coldness of the freezing water, and endless related memories and emotions. Just like the score is a representation that allows someone to play a piece of music with an instrument, the activation of neurons corresponds to a representation of some concept that, in turn, offers a pathway to the activation of related representations making us *feel* things.

And so, we must conclude that it is not correct to view dreams as an innocuous consequence—an epiphenomenon—of cerebral activations. Neural activations and what they represent are one and the same thing, though we still have to resolve whether random activations during dreams have a specific function.

One Hundred Years of Solitude, by the brilliant Gabriel García Márquez, is likely the most renowned novel written in Spanish. Published in 1967, it marks the pinnacle of magic realism, a literary movement developed by Latin American writers in the twentieth century, where they masterfully interweave reality with fantasy. García Márquez's imagination had no limits and goes from conceiving of a woman who scratches the walls to eat earth and limestone, to an almost-150-year-old woman who is buried in her rocking chair, a young man who is always followed by yellow butterflies, or a baby with a pig's tail who is carted off by ants.

The action takes place in Macondo, a fictitious town that is home to Aureliano Buendía and his large family. Daily routines go on without any major disturbances, until one day Macondo is hit with the "plague of insomnia" (and what better example of magic realism!). The epidemic quickly spreads among residents, who at first are happy to have more time to finish everything they couldn't do during the day, but soon after, they discover the monumental boredom that comes with being constantly awake, and an even more dramatic consequence: oblivion.* As the nights of insomnia go by, memory loss becomes more apparent. To try to overcome the situation, Aureliano Buendía begins to hang small signs with the names of things: anvil, table, chair, clock, door. However, with oblivion's accelerated progress, the name alone is not enough and, for example, he hangs a cow sign that says: *This is the cow. She must be milked every morning so that she can produce milk, and the milk must be boiled to combine it with coffee to make coffee with milk.* José Arcadio Buendía, his father, decides to build a "memory machine" with several thousand cards that flip by one after another on a rotating device, and every morning they remind him of the knowledge he has acquired throughout his life. The plague of insomnia wreaks havoc in Macondo until one day a gypsy called Melquíades brings along a memory-healing potion.[**]

We have all heard it is a bad idea to pull an all-nighter studying for an exam. Aside from being tired during the test, the problem is that we will not have assimilated the new information because we didn't sleep. To confirm this, in the early 1920s, John Jenkins and Karl Dallenbach tested the role sleep has on memory, comparing the number of senseless syllables a group of subjects could remember at different times (having slept and not). As expected, they found that the number of syllables people remembered decreased with time—the more time that went by, the less the subjects remembered—but Jenkins and Dallenbach also noticed that the subjects' performance noticeably improved if they had slept before being evaluated.[16]

We can agree that sleep plays a key role in memory, but can we say the same about dreams? What is it that really matters: the rest we get while we sleep or the actual construction of dreams? Deep down, we wonder if dreams are really "just froth," as gleaned from Hobson and McCarley's model, or if they have some function, though perhaps not the one proposed by Freud or Jung. To answer this question, we can use the fact that dreams usually take place during the REM stage and compare the effects that both REM and NREM stages have on the consolidation of explicit memory, meaning our memory of facts and experiences.

Except in pathologies like narcolepsy, we saw that sleep stages follow a structure that is repeated approximately every ninety minutes. The beginning of sleep is dominated by the NREM stage, and the first appearance of the REM stage usually happens after an hour of sleeping. Therefore, a typical short nap, of less than an hour, lacks REM sleep. A group of German researchers used this principle to evaluate the influence of NREM sleep in memory consolidation. They gave a group of subjects a list of words to remember and then evaluated how many words the subjects could remember in a short time, both if they remained awake or if they took a short nap. The surprising result was that even after only a six-minute nap, the subjects remembered more words than if they remained awake.[17] In line with this result, another group of German researchers noticed that memory consolidation was stronger in subjects who only slept during the first hours of the night (from 11 PM to 2 AM), dominated by NREM sleep, compared to those who slept during later hours (3 AM to 6 AM), dominated by REM sleep.[18]

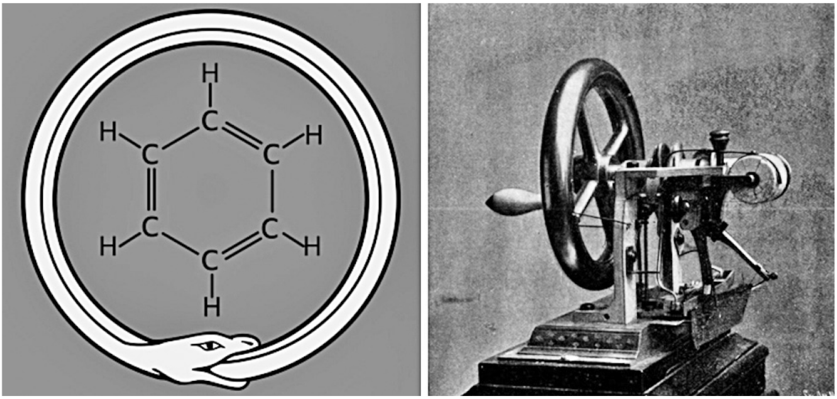
This and several other experiments clearly prove that NREM sleep is vital for memory consolidation.[19] However, a proof of REM sleep's role (and particularly the role of dreams) in memory is relatively scarce. To tackle this question, Robert Stickgold—a researcher at Harvard University and one of the most renowned experts in sleep study—evaluated how a group of subjects resolved anagrams or established semantic connections among words immediately after having been woken up from REM and NREM stages of sleep.[20]

Usually upon waking up we remain drowsy for a few minutes. Likewise, when awakening in the REM or NREM stage, the brain remains in the state it was in for a few minutes before completely waking up. Stickgold's trick was to carry out experiments only a couple of minutes after waking up the subjects, therefore assuming that the brain was continuing to process information as if it were still in the preceding sleep stage. The result was that the subjects' performance was much better when they were woken from the REM stage, the one where most dreams are produced. Stickgold concluded that the REM stage, and particularly dreaming, favors a greater state of fluidity of thought compared to the NREM stage.

Dreams are influenced by the situations we experience during the day, although they are far from being a repetition of those experiences. To show this, Stickgold and his colleagues asked a group of subjects about their dreams and experiences during the prior few days, and they observed that although dreams touched on the general themes of these experiences (like in Deirdre Barrett's previously mentioned experiments), only 1 or 2 percent of them were literally repeating the experiences.[21]

As we saw earlier, the consolidation of experiences in our memory does not happen while we dream but during the NREM stage, when we dream less. Stickgold argues, in my opinion to great effect, that REM sleep has a much more interesting function: putting these experiences in context, comparing and associating them with the rest of our knowledge. It is worth noting that the associations made during sleep are not limited to the immediate experiences during wakefulness, but they cover a wide range of stored knowledge. Attempting such disparate associations, rather than those involving things that are coherent and closely related to one another, makes dreams so bizarre because the random activations described by Hobson and McCarley test the connection between new memories and those that have already been acquired, creating a tangled mess of associations that shape our memory and knowledge. In conclusion, the NREM stage reinforces new memories, while the REM stage and dreams put these memories into context, which causes us to not only remember but also understand.[22]

The associations we test and establish during our dreams are also the cornerstone of our creativity. As we saw at the end of chapter two, creativity is based on establishing associations among disparate concepts. Newton's genius was realizing that an apple that falls and the moon that orbits around us respond to the same phenomenon: the law of gravity. And precisely these types of disparate associations are the ones that happen in dreams. For example, German chemist August Kekulé discovered benzene's cyclic structure after dreaming of a snake biting its tail. American inventor Elias Howe discovered the working principle of the sewing machine after dreaming he was surrounded by a group of aborigines with spears. Robert Louis Stevenson imagined part of his famous novel, *Dr. Jekyll and Mr. Hyde*, during a dream, just like Mary Shelley conceived *Frankenstein* in a nightmare. Closer in time, Paul McCartney, of Beatles fame, dreamed the melody of "Yesterday" and wrote it down when he woke up.



Benzene structure, discovered by August Kekulé (1829–1896), after dreaming of a snake biting its tail, and a sewing machine designed by Elias Howe (1819–1867), after dreaming of a group of tribesmen surrounding him with spears.

Although scientific evidence is still relatively scarce, we have seen that, far from being an epiphenomenon, dreams apparently process, contextualize, and assimilate information related to what we have experienced during the day. The ability to manipulate other people's dreams (or our own) as in the movie *Inception* is relatively limited, and likely for the best, since if we managed to manipulate dreams, we might end up affecting the key function they have in the workings of our brain. However, we'll see that one of the movie's predictions is much more real and concrete. Instead of limiting themselves to digging around people's dreams—which is already a lot—Nolan's characters decide to try something even more daring: to implant an idea, a memory. In the previous chapter, we argued that our memory defines who we are. So, is it possible to implant a memory? And if so, what consequences would this have on our personal identity? This is exactly what we will explore in the next chapter.

---

* This brings to mind Zeno's famous paradox of the tortoise and Achilles: The legendary Greek warrior races against a tortoise, but the tortoise is granted a head start, because it is slower. When Achilles reaches the tortoise's starting point, it has already gained some ground; when Achilles reaches the following spot, the tortoise is once again somewhat ahead of him, and this goes on and on, with Achilles apparently never reaching the tortoise. This paradox—like the idea of prolonging someone's life by immersing him in a deeper dream that slows time—is based on the fact that there are always endless points within an interval in time, no matter how small it may be.

* On the other hand, given the infringements of the laws of physics and time perception within dreams, Cobb's top continues spinning indefinitely when he is not awake.

* Adapted from: A. Rogers, M. Aldrich, and C. Caruso, "Patterns of Sleep and Wakefulness in Treated Narcolepsy Subjects," *Sleep* 17, no. 7 (1994): 590–597.

* To be able to ask ourselves these questions during a dream, we must ask these same questions while we're awake. As ridiculous as it might sound, for example, to confirm that we cannot penetrate a wall with our hand, by trying this repeatedly while awake, we increase our chances of eventually being able to do the same thing while asleep.

* The eye acts as a dipole, with a positive charge at the front and a negative at the back, that generates an electric field around it. The electrooculogram measures variations of this electric field with electrodes placed around the eye, in order to estimate eye movements.

** Figure adapted from Stephen LaBerge, "Lucid Dreaming: Psychophysiological Studies of Consciousness During REM Sleep," in Richard R. Bootzin, John F. Kihlstrom, and Daniel L. Schacter (eds.), *Sleep and Cognition* (Washington, DC: APA Press, 1990), 109–126.

* Later, the NREM stage was subdivided into four sleep phases, according to stereotyped patterns in the EEG recordings: phase 1, typical when falling asleep, is when we are sleeping lightly and can be easily awakened; phase 2 is usually the transition from or toward deep sleep in phases 3 and 4 (when it is hardest for us to wake up).

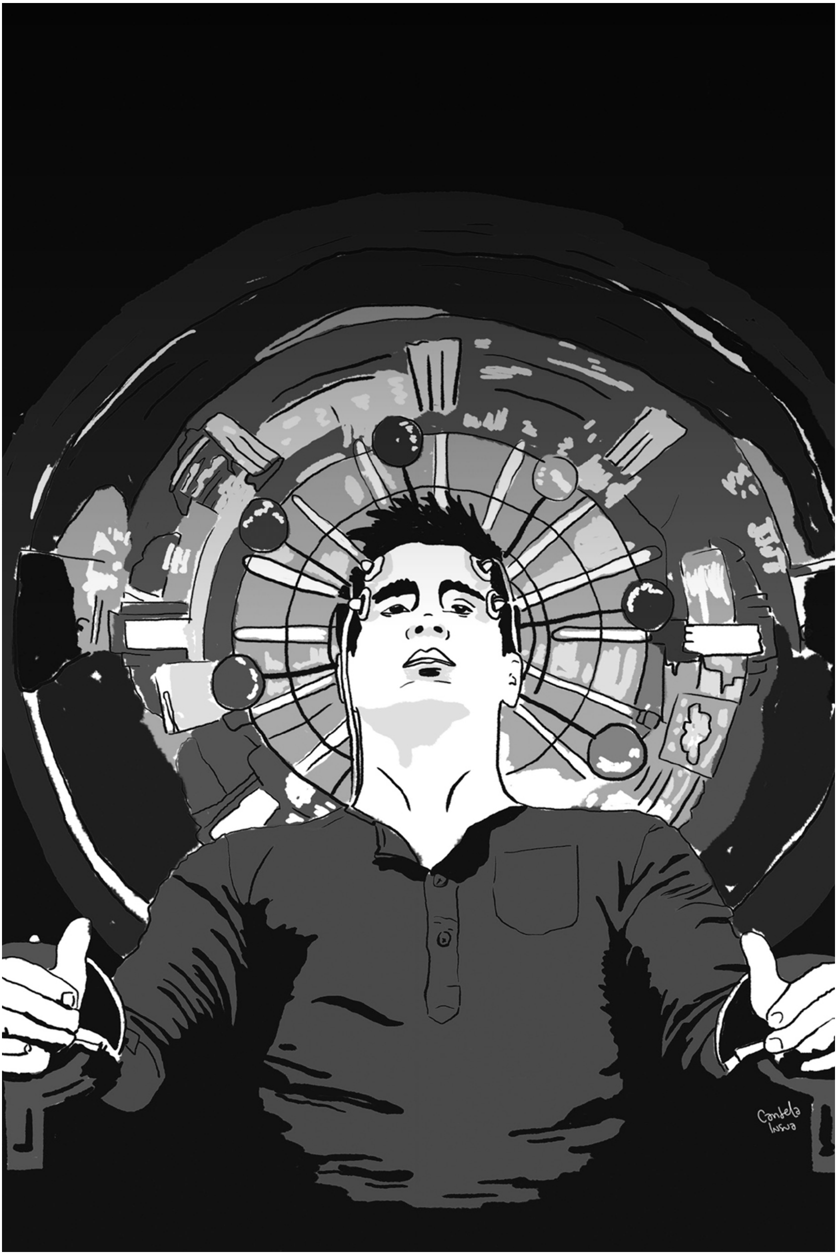
* Prolonged insomnia not only has catastrophic memory consequences, but it also impacts the person's health. It occurs in an extremely rare and incurable disease called *fatal familial insomnia*. Patients with this condition begin to noticeably deteriorate after a few weeks without sleep (similar to patients with advanced dementia) and die a few months later.

** In a tragic irony of fate, García Márquez had Alzheimer's during the last few years of his life. Unfortunately, we still do not have a magic potion like that of Melquíades to cure the devastating effects of the real "disease of oblivion."

CHAPTER 9

TOTAL RECALL

Memory Manipulation and Implantation


Cantela
Insua

Douglas Quaid is a construction worker who dreams of going to Mars. Desirous to escape the monotonous routine of his life, and tempted by a subway ad, he decides to visit the offices of "Rekall" to take a virtual trip to the Red Planet via a memory implant. The Rekall salesperson mentions several advantages of taking a virtual trip compared to physically traveling to Mars: It is cheaper, safer, with first-class travel, a deluxe hotel, and none of the complications of real-life vacations. Quaid doubts and wonders how credible the journey's experiences will be, but the employee assures him that the vacation will be just as real as any other memory in his head because his brain won't notice the difference.

While arranging the details for his trip, Quaid automatically rejects the additional options offered until the employee asks him to name the one and only thing that remains exactly the same in each of his vacations. Quaid doesn't have an answer, and the salesman promptly reveals that what is always a constant is Quaid himself, because he is still the same person regardless of the place or circumstances. The option he ends up finding irresistible is the "Ego Trip," where he can take on an alternate identity. Ultimately, it is all a memory implant, and there's nothing stopping Quaid from being an industrial tycoon, a sports hero, or a secret agent, which is what he quickly chooses.

Everything is ready for Quaid to enjoy his journey to Mars, but once the procedure is underway, something goes wrong. He begins to recover lost memories and eventually discovers his name is not Quaid but Hauser, a secret agent on Mars, and that both his identity and his past, as well as his memories of being married for several years to a woman named Lori, are just memory implants. He finally manages to travel to Mars, where he is ensnared in a monumental plot that leaves him little choice but to join the resistance against an evil government corporation that rations and charges for the use of breathable air.

After several chases and shoot-outs, a doctor shows up at Quaid's hotel and tells him that he's actually not on Mars but in an "implant chair" at Rekall, experiencing a memory implant that is getting out of hand, and that he has been artificially introduced into Quaid's mind to save him from what is turning into a nightmare. This makes sense, since Quaid's life took a dramatic turn right after starting his virtual vacation. However, it also makes sense that the Mars vacation implant had awakened partially erased memories, and that his desire to see Mars was nothing more than memories surfacing from his subconscious.

This is basically the dilemma presented in *Total Recall*.* Who is the person facing these puzzling situations? Is it Quaid, a simple construction worker experiencing the virtual reality of being a secret agent, or is it Hauser, a secret agent who has had Quaid's identity artificially implanted in him to erase his past? The ambiguity of not being able to choose between his two identities dominates the rest of the plot and the possible interpretations of the movie.

When Quaid is trapped by the villains, their leader (the corrupt governor of Mars and Quaid's main antagonist) shows him a video of himself in the past with Hauser's personality. In the video, Hauser confirms that Quaid's current identity is a memory implant that he—the Hauser of the past—is about to receive. He also says that when the time comes, he will want to recover his original memories and become himself again. Quaid resists, because deep down he despises Hauser—someone who has his same body but is not him because he now has another set of memories.

Let's go back to Borges one more time. In "Shakespeare's Memory," he narrates the misfortunes of Hermann Sörgel, a World War I–era literature professor to whom a colleague, after a long night at an English pub, offers something unique: William Shakespeare's memory. Sörgel accepts, somewhat skeptical, yet also longing to share the mind of the most brilliant writer of all time. He slowly begins to access the Bard's private thoughts, yet, after his initial fascination, ends up rejecting Shakespeare's memories after realizing that he is starting to confuse them with his own and slowly beginning to lose his identity. Says Sörgel in Borges's story:

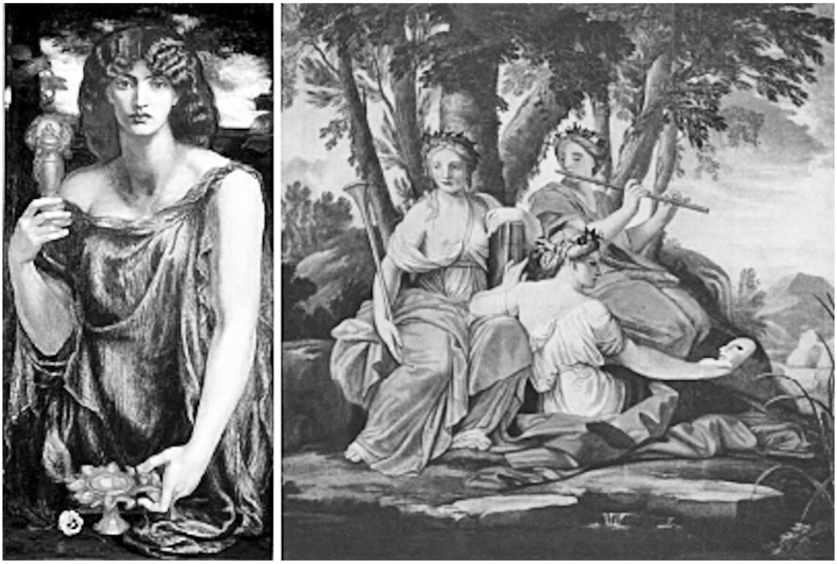
> In the adventure's first stage I felt the fortune of being Shakespeare; in the final one, oppression and terror. At the beginning, both memories did not merge their waters. With time, Shakespeare's grand river threatened, and almost flooded, my modest stream. I fearfully realized I was forgetting my parents' language. Since personal identity is based on memory, I feared for my reasoning.
>
> —"*La memoria de Shakespeare*"*

We saw in chapter seven that in the seventeenth century, John Locke used the paradox of the prince and the cobbler to pose the question of what determines identity. Locke, like Borges (or *Total Recall*'s plot), asserts that identity is tied to memory. Memories make up a person. As for myself, I am the one who studied at the University of Buenos Aires, then pursued a PhD at Lübeck, and am now writing this book in England. But how reliable are the memories that make me who I am? Did it really take eight hours to get through that classical mechanics exam, where we brought a snack to avoid passing out, while we tried to solve seemingly endless problems? Did the idea to write this book really come from that talk I gave at the Colegio Nacional de Buenos Aires?

If there's something I'm certain about, it is my own existence, my past, and knowing who I am. But I mentioned in chapter four how misleading one of my childhood memories was, where I misremembered the orange ball at the winter soccer match in Kiev. I then wonder just how fragile and malleable is my memory and, consequently, my identity. I wonder, likewise, if someone could implant memories in my brain and change my personality, like in *Total Recall*.

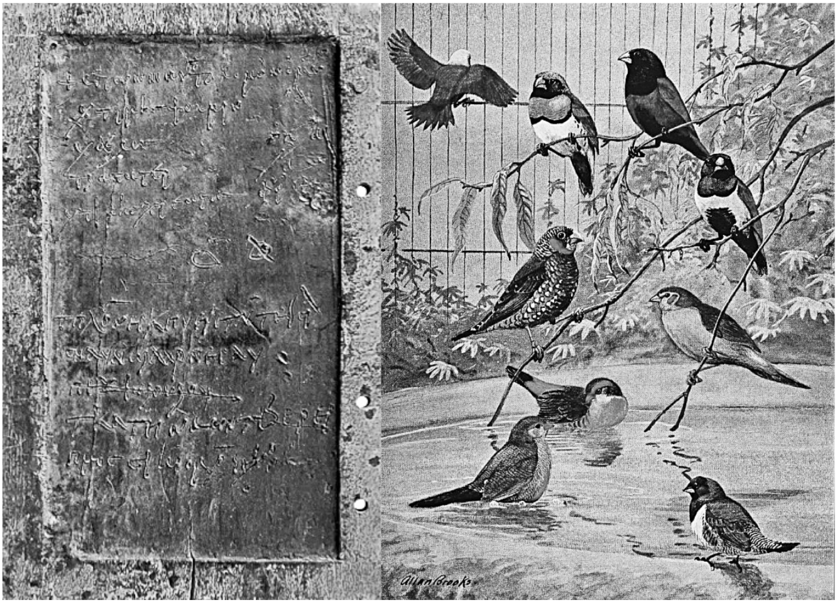
In Greek mythology, memory is personified in the goddess Mnemosyne, the mother of the nine muses.* But why have a memory goddess? It is understandable to have a goddess of the hunt (Artemis), of knowledge (Athena), of love and beauty (Aphrodite), or of the earth (Gaia). But memory? Why not feelings, impulses, or consciousness?

One of Plato's dialogues describes how Crito invokes Mnemosyne before a long presentation, because he fears not remembering what he has to say.[1] We cannot underestimate the importance of memory in ancient times, since, lacking paper, speakers depended exclusively on their memory to give a speech; so much so, that two of the three fundamental period books about memory refer to public speaking: *De Oratore*, by Cicero, and *Institutio Oratoria*, by Quintilian (the third one is *Rhetorica ad Herennium*, of unknown authorship).



Mnemosyne (oil painting by Dante Gabriel Rossetti) and three of the nine muses (oil painting by Eustache Le Sueur): Clio (history), Euterpe (music), and Thalia (comedy).

Paper manufacturing began during the Han dynasty, in the early second century in China, but it would take more than a millennium to reach the West. The technology spread to the Muslim world during the Battle of Talas (751 AD) between the Abbasid Caliphate and the Chinese Tang dynasty, where prisoners of war showed the Arabs how to process paper in exchange for their freedom. The technique for making paper only reached Europe in the twelfth century, after the reconquest of southern Spain. Until then, scriptures were written on papyrus—which was very costly, because it was imported from Egypt—or on the even more expensive parchment, made from animal skin. In ancient times, particularly in Ancient Greece, inscriptions on wax tablets (wooden tablets covered with wax) were also used, but these writing techniques were very cumbersome and costly, and therefore of limited use.



Plato compared memory to the inscription on a wax tablet or a birdcage that is gradually filled with birds.*

Plato describes memory as the printed mark on a block of wax, saying that when the wax in the mind of a person is pure, deep, and abundant, the memories have enough space to be established and endure.[2] In the same treatise, he considers memory as a birdcage. The cage is empty in a child and as he grows, he begins to fill it with birds, representing each memory or piece of knowledge. In this way, Plato differentiates between having knowledge and accessing it (what nowadays is called *encoding* and *recalling* a memory). To have knowledge is to have a cage filled with birds, while accessing knowledge implies being able to grab any of these birds, examine it, and place it back in the cage. Plato's two images reflect the intuitive idea we have about memory, though they differ quite a bit from reality. Far from being fixed etchings on wax or birds in a cage waiting to be put to use, the formation and storage of memories is a much more dynamic process.

The experimental and systematic study of memory began with German psychologist Hermann Ebbinghaus, who at the end of the nineteenth century carried out a series of simple yet tedious experiments on himself. He constructed 2,300 senseless words, each one made up of three letters (two consonants with a vowel in the middle), and in endless experiments, he evaluated how many of these randomly chosen words he remembered after different time intervals.[3]

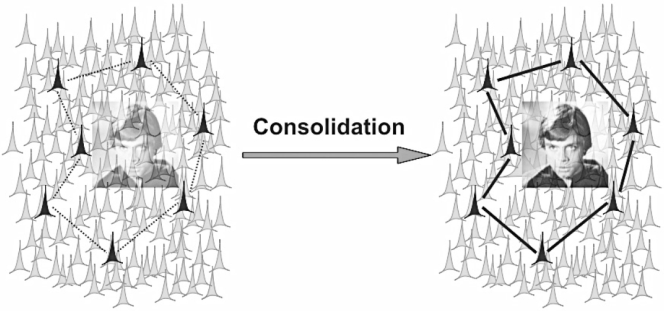
Ebbinghaus determined two fundamental principles. First, while some memories don't last more than a few minutes, others persist for hours, weeks, or even years; this is what is currently known as short- and long-term memory. *Short-term memory* keeps information for brief time intervals and makes us aware of the present sequence of events. It is the one I use, for example, to remember what I want to say in the moment, as I choose the right words. In contrast, *long-term memory* is made up of things we remember from our past, for example, in my case, the memory of a good wine, my mother's face, or a day at the beach.

Almost all short-term memories are lost in oblivion and only a minuscule part of what is happening to us right now will become part of our long-term memory. But what makes some memories fleeting and others last a lifetime? This is Ebbinghaus's second principle: Repetition and practice make memories endure, because when reviewing his results, he noticed that the more he repeated a list of words, the longer he remembered them. Long-term memory is specifically made up of events that stand out and to which we pay attention; these are memories we keep bringing back into our consciousness and that will consolidate with each repetition.

Using experiments similar to those done by Ebbinghaus, German psychologists Georg Müller and Alfons Pilzecker observed at the end of the nineteenth century that learning a list of words was interrupted if they asked subjects to learn another list immediately after the first, although not so much if they asked them to learn the second list a while later. With these results, they concluded that there's a critical period during which learning the second list interferes with the consolidation of the first. This period of consolidation explains the temporary amnesia experienced by boxers after a knockout, where at times they don't even remember going to the fight. Amnesia happens due to a loss of consciousness caused by the received blow, which interrupts the consolidation of previous memories. However, beyond the existence of a critical period during which memories become fixed, consolidation is a constant process. For example, we tend to have more vivid memories of events for which we have photographs that we look at once in a while. These photos make us review the events over and over again and consolidate them in our memory.

For the last one hundred years, the consolidation model has been a cornerstone of our understanding of how memory works: When recently acquired, memories are unstable and represented by specific connections between neurons that, if consolidated (for example, through repetition), will make the memory endure in time.[4] Following Plato's analogy, if memories are like wax etchings, the more we recall a memory, the deeper the etching will be.

So, there seems to be a very small margin for deleting or implanting memories like in *Total Recall*, since, according to consolidation theory, the only moment when we could manipulate a memory is right after it has been acquired, because once it is consolidated it will last for the person's lifetime. However, in 1968, a group of American scientists published a series of results that didn't fit this viewpoint.[5]



Classic memory consolidation model. If we see *Star Wars* icon Luke Skywalker for the first time, we initially form an unstable memory of him, represented by a group of interconnected neurons. When repeating and consolidating the memory of Luke, the connections become stronger and strengthen the memory.

We all know about Pavlov's conditioned responses: If a dog is given food right after a bell sounds, the dog associates the bell with the food, and by sounding the bell alone, the canine begins to salivate, awaiting its meal. A variation of this experiment is *fear conditioning*, often used in rats because they learn it almost immediately. In this case, a specific sound is presented, followed by a small electric shock that makes the animal uncomfortable, and, once this has been repeated a few times, it is remarkable how after hearing the sound, the rat prepares itself to receive the shock. On the other hand, we have also likely heard about electroshock, which consists of applying a strong electric current to treat certain pathologies, and it has an effect similar to a knockout punch, creating a temporary amnesia.

Bringing these two pieces of information together, the experiment that demolished the classic consolidation theory was very simple. First, researchers created a conditioned fear response in rats: A sound was followed by a small, yet uncomfortable electric shock. The animals quickly learned to associate both stimuli and adjusted their behavior, drinking much less water due to their state of anxiety. Immediately after the conditioned response had been created, a group of these rats received a big electroshock to interfere with the consolidation of this memory, and when the experiment was repeated the following day, these rats didn't remember the previous day's learning session and drank water normally. Even more interesting is what the team did with a third group of rats, to which they applied the electroshock not after the first day of learning but after going through the same task a day later; in other words, and this is the crux of the experiment, once the rats had learned and consolidated the conditioned response. The surprising result was that this made them forget what they had learned the previous day, and they again drank water normally. In other words, by reactivating and interfering with an already consolidated memory, they managed to erase it.

Consolidation theory was so prominent during the last century that these results went relatively unnoticed until a similar experiment, using a pharmacological manipulation in rats, reached the same conclusion thirty years later.[6] We now know that consolidation is a dynamic process. There are no static etchings in wax; memories change in a constant *reconsolidation* process.[7] Therefore, memories are unstable not only when they are acquired (as stated by consolidation theory) but also each time they are recalled. The act of accessing a memory implies changing it.

The idea of reconsolidation, of being able to modify or even erase certain memories, has extraordinary therapeutic connotations. Many people suffer *post-traumatic stress disorder* caused by traumatic experiences such as sexual assault, a car accident, or extreme situations experienced at war. People suffering this disorder cannot forget those experiences, and when involuntarily recalled, the unwanted memories create stress and panic. The challenge is that these patients typically only reach out for professional help when such stressful reactions have already been consolidated in the brain, and ideally one would like to intervene in the consolidation of unwanted memories as soon as they happen, which is obviously impossible to implement in practice. However, if memories enter an unstable state each time they are consciously brought forward, we could evoke an unwanted memory and then try to erase it.

Implementing the previously described experiments on humans isn't so easy because, on the one hand, electroshock is only used in extreme cases as it has significant contraindications, and, on the other, the drugs that were used in studies with animals are highly toxic, though less toxic drugs could have the same effect.

Propranolol is a blood pressure medicine taken by millions of patients every day. It is what is known as a *beta-blocker*; its main function is to decrease the heart rate. However, the drug also has the side effect of blocking the synthesis of proteins involved in maintaining connections between neurons, and, as we previously illustrated with Luke Skywalker, the connectivity between neurons is the basis of memory. Merel Kindt, a Dutch researcher, tried to use propranolol to interrupt memory reconsolidation in humans.[8] The experiment was based on establishing a conditioned response (the presentation of a picture, followed by an uncomfortable electric shock), then evoking the response the following day, after giving the subjects propranolol, and finally reviewing the reactions to the pictures on the third day. Kindt and her colleagues observed that the application of propranolol effectively managed to erase the conditioned responses.

Despite these promising results, there's a huge difference between eliminating the reaction to a small electric shock and erasing a traumatic memory from the past. While some studies state that it is possible to relieve post-traumatic stress disorders by getting the patients to relive those experiences while taking propranolol, others conclude that the clinical use of this procedure is relatively limited.[9] It is important to note that even in the best of cases, the reconsolidation of traumatic memories does not erase the memory itself but rather the subject's emotional reaction. In other words, war veterans will continue remembering for the rest of their lives the dramatic situations they have experienced, but the idea is for them to do so without overstressing themselves or panicking. Completely eliminating a memory is much more controversial from an ethical standpoint, since our memories, no matter how dramatic, are part of our lives and personalities: the learning we turn to when trying to avoid repeating unhealthy patterns or circumstances associated with a bad experience.

It is usually almost impossible to completely eliminate a deeply rooted experience like a traumatic memory (again, the previously mentioned experiments only eliminated the emotional reaction the memories generated), but we could probably alter memories that are vaguer and more unstable, and without the use of any drug, just through suggestion. The great expert on this subject is psychologist Elizabeth Loftus, who in a famous experiment showed how memory reconsolidation can lead to fabulation and the creation of false memories.[10]

Loftus showed different subjects a video of a traffic accident and later asked them to estimate the speed of the cars. The interesting twist was that while she asked one group of subjects about the cars' speeds when they *bumped* into each other, she asked another group the same question but using the word *collided*; with a third group she used the word *hit*; with a fourth group she used the word *contacted*; and with the last group, *smashed*. Every subject saw the exact same video and under the same conditions, but those with whom she used the word *smashed* gave the highest speed estimate, followed by those with whom she used *collided*, then *hit*, *bumped*, and finally *contacted*. Even more surprising was the fact that a week later she asked these same people if they remembered having seen broken windows at the scene of the accident and, while 32 percent of the subjects with whom she used the word *smashed* responded (incorrectly) yes, only 14 percent with whom she used the word *bumped* stated they had seen them.

Loftus's results show how fragile our memories are and how they could be manipulated in the reconsolidation process by simply changing one word in one question.[*] However, beyond changing the impression of something we have seen or experienced, could we use suggestion to implant a memory of something that has never happened?

Following this line of experimentation, Loftus gave a group of people descriptions of four events from their childhoods, provided by their relatives. What the subjects didn't know was that one of these memories—having been lost at a shopping mall—was completely false. Loftus asked the subjects to write what they remembered about these stories during five consecutive days, and although at first they doubted the event at the mall, they ended up being convinced that it really happened. After each recap, one of these subjects, for example, claimed to remember more and more details of that day: the kind, balding man with a blue shirt and glasses who had helped him; the fear he felt about not seeing his family again; his mother scolding him when she found him—all false memories. Finally, he was told that one of the four childhood memories was fake and when asked to guess which one, he chose a real memory, in complete disbelief that he was never lost in a shopping mall. Likewise, Loftus describes how a few shameless psychoanalysts used suggestion and reconsolidation to make patients believe they had suffered sexual abuse and other dramatic events in their childhood (these patients later won millions in lawsuits after proving that these memories were nothing more than lies implanted in their memories).[11]

We previously discussed how memory is a creative process, and the fact that we actually remember very little information and reconstruct our memories using inferences, generating Bartlett schemas, the foundations of our memories. Now we have seen that this construction is a very dynamic process that changes with time, and it allows the manipulation and even the implantation of memories.

Beyond the undisputed merit of scientists such as Elizabeth Loftus, I believe the great conjurers, the great memory manipulators, are magicians, who with their art lead us to experience the impossible. For example, magicians can recap what has happened before unveiling a trick, but this recap tends to be partial because they reinforce some facts they want people to remember and ignore others that they prefer to remain forgotten, since the latter could give the audience clues about how the trick was performed. The seasoned magician will not only influence but also alter the content of these memories as they are reconsolidated. He could, for example, quickly shuffle and hand a stack of cards to a spectator who will cut it several times (which has no consequence whatsoever). After numerous passes and distractions, he could summarize what has happened by saying something like “the cards have been shuffled,” looking at the spectator while that person holds the deck. This is the false memory seed; a vague yet correct affirmation, since it doesn’t specify who shuffled the cards, although it suggests the idea that it is the spectator who did the shuffling. If instead of a mildly misleading suggestion, the magician would use a more precise affirmation like “the spectator has shuffled the cards,” the audience would immediately notice the deception. The magician’s talent resides in reinforcing these ideas over and over again—but without explicitly saying so—to end up implanting a memory that leads to something impossible, for example, that he could find a card in a stack he hasn’t manipulated.[12]

Miguel Ángel Gea performing close-up magic. The art of the magician relies not only on dexterity at manipulating cards but also on shifting the attention of the audience and their memory of what he does in his tricks.

Half a century ago, in the faraway Ural mountains lived Akakhi Akakhievitch, a brilliant yet unknown Russian neurosurgeon. Convinced that ideas are represented by specific neurons, he began to experiment on animals (and on Trotskyists awaiting their death sentences in Siberia), trying to locate the ones that coded the concept "mother." After several attempts, he found neurons responding only to the mother of the subjects, regardless of how she was shown: face forward or in profile, upside down, as a cartoon, and so on. As fate would have it, Akakhievitch then received a patient who begged him to help him forget his authoritarian and unbearable mother. Using his recent knowledge, the surgeon proceeded to extract each and every one of the thousands of neurons that represented the concept "mother" in the patient's brain. When he awoke from the anesthesia, the man had in effect lost the idea of his mother; every memory related to her had vanished. Elated by the successful surgery, Akakhievitch decided to face the following challenge: to find the "grandmother cells."

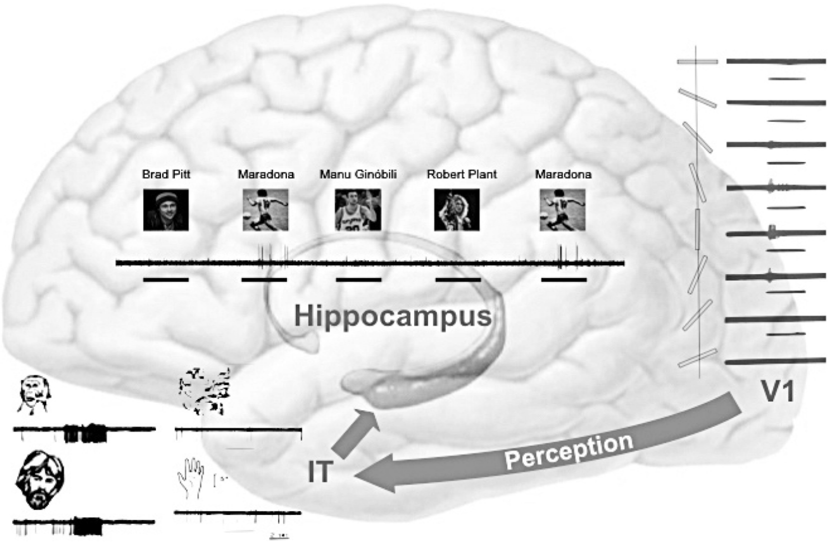
The story, clearly a hoax, was conceived by neuroscientist Jerry Lettvin to spice up a class he taught at Massachusetts Institute of Technology in the late 1960s.[13] Taken as a joke at first, Akakhievitch's story ended up setting off an intense debate in neuroscience. But the discussion of the existence of neurons representing specific things is much older and goes back to William James, the father of modern psychology, who in the late nineteenth century proposed that "among all cells there may be a central or pontifical one linked to our consciousness."*

The idea of *pontifical cells* would go unnoticed for decades, but it was picked up by Nobel Prize winner Charles Sherrington, who argued that in order to decide if, for example, we are seeing the face of a friend, our mother, or an apple, the brain does not use a single pontifical cell—meaning unique neurons—but a millionfold democracy.** In other words, according to Sherrington, the representation of concepts is not due to the activation of specific neurons but instead distributed across large populations of neurons. These neurons respond to different minor details, and by adding all the details (for example, the nose, skin tone, eyes, hair, lips, and so on), a representation of the concept is generated.

To scientifically prove the possibility of the existence or not of "grandmother cells," we need to talk a bit about visual perception. The information captured by our eyes is transmitted to the primary visual cortex, where David Hubel and Torsten Wiesel discovered, in the 1950s, neurons that respond to lines with different orientations, results that would garner them the Nobel Prize in 1981. This information reaches the inferior temporal lobe (IT), where, through experiments with monkeys, American scientist Charles Gross found neurons that specifically respond to faces. (In about the same area, Doris Tsao and Winrich Freiwald later discovered the face patches we discussed in chapter five).

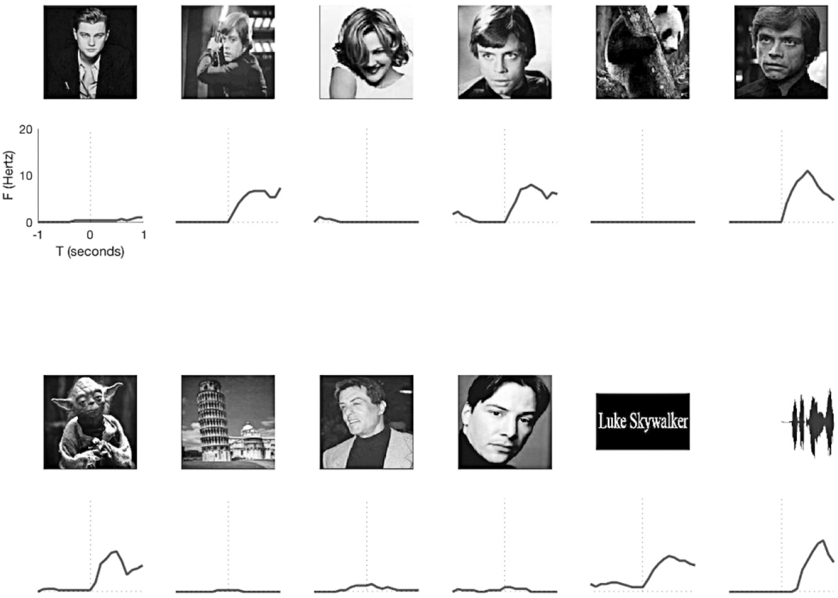
As we proceed in the processing of visual information in different areas of our brains, the information represented by each neuron becomes more complex. Taking this hierarchal processing into account, in the late 1960s, Polish psychologist Jerzy Konorski wondered if this configuration wouldn't end up in individual neurons representing specific objects or persons (as proposed by James decades earlier and ridiculed by Lettvin a few years later with his story of Akakhi Akakhievitch).[14] In fact, after a visit to Konorski's lab, Charles Gross thought of looking in high-level visual areas for responses to faces. But the neurons discovered by Gross are far from the ones Konorski predicted, because they respond to faces in general and not specific individuals.

The neurons in high-level visual areas project to the hippocampus, where in previous chapters I mentioned that—experimenting on patients with epilepsy, implanted with intracranial electrodes for clinical reasons—we found neurons that do respond to specific people, like Jennifer Aniston, Halle Berry, or Luke Skywalker. In other words, these neurons don't respond to a particular picture of a person, but to the concept. For example, the Luke Skywalker neuron responded to three different pictures of the character and even to his name in writing or when spoken by a computer-synthesized voice. So far, that's very close to Lettvin's "grandmother cells"—cells that respond to one person regardless of how that person is seen—but we must consider this claim more carefully.



Processing of visual information. Neurons in the primary visual cortex (V1, in the back of the brain) respond to lines with specific orientations (in this case a vertical line). This information is transmitted to high-level visual areas (through the so-called perception pathway) and end up in the inferior temporal cortex (IT), where neurons preferentially responding to faces have been found. The information in the IT is directly transmitted to the hippocampus, where we found neurons that respond to specific concepts—in this case, Maradona.[15]

I'm delving into the details of how visual information is processed here because they are relevant when considering the possibility of artificially erasing or manipulating memories and, more broadly, to understanding how memory works. Above all, if we once found a neuron responding to Jennifer Aniston, this is one of many, because the possibility of finding the one and only neuron (among billions) responding to a specific concept is practically zero. In other words, if we found a neuron responding to Aniston, there have to be more, forming what is known as a *cell assembly* or *engram*, a group of interconnected neurons that represent a memory. Furthermore, even if accepting that there are perhaps thousands of neurons responding to the actress, it is impossible to prove that these neurons are only responding to Aniston and nothing else. To prove this, we would have to show pictures of every person the subjects knew, which is impossible to do. Furthermore, the day after we found the Jennifer Aniston neuron, we repeated the experiment including pictures of other actors from *Friends*, the TV series that made her famous. The neuron once again responded to Aniston, like on the previous day, but also responded to Lisa Kudrow, who acted in the same series. Likewise, the neuron that became active with Luke Skywalker also responded to a picture of Yoda—another *Star Wars* character—and throughout the years, we have observed countless examples where these neurons respond to more than one concept; although if this is the case, the concepts they respond to are usually related.[16]

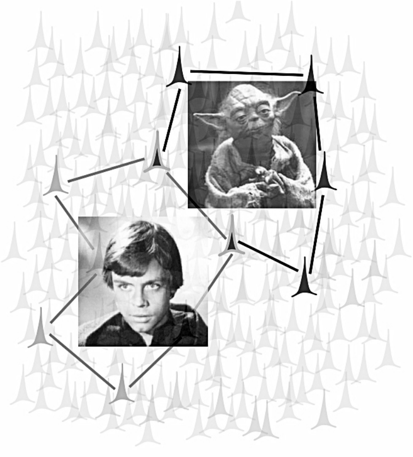


Responses of one neuron to three photos of Luke Skywalker and his name written and spoken by the computer (panels in the bottom right corner). The neuron also responded to Yoda, another *Star Wars* character.

We have already seen that concept cells are found in the hippocampus, a key memory area, which makes sense because we tend to remember concepts—the meaning of things—and forget details. I remember, for example, a day with my brother at a beach in Brazil, but I don’t remember how he looked that day or the color of his shirt or swimsuit. All those details have been lost in oblivion, whereas we tend to remember concepts: brother, beach, Brazil. But now we must add another piece of information to the puzzle: These neurons not only code concepts, but also associations among them, since they tend to respond to related concepts. But why? Why does one neuron that is activated by different pictures of Luke Skywalker also respond to Yoda? Why does the Jennifer Aniston neuron also respond to Lisa Kudrow?

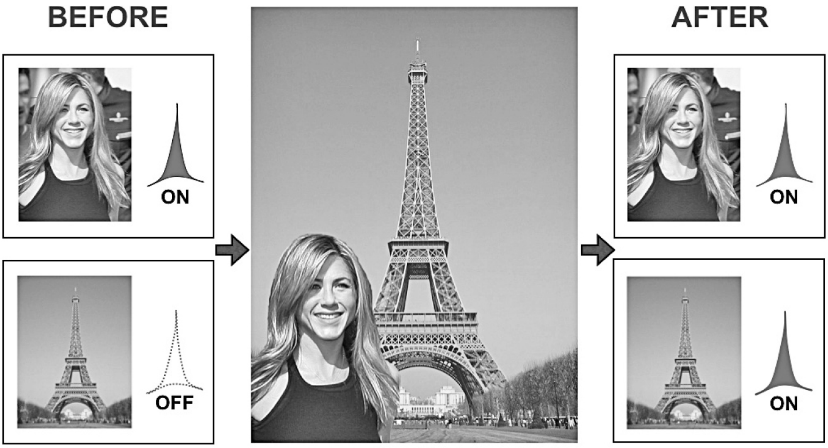
William James, and much earlier Aristotle, in *De memoria et reminiscentia*,[*] argued that associations constitute a fundamental memory mechanism. To think is to make abstractions, associate concepts, and create relationships and categories. These associations are precisely the ones that help us strengthen and evoke memories. By generating associations and context, we may perhaps not remember a specific fact, but we will likely remember some associated facts, and through these, we can reach the one we were looking for. However, associations are much more than memory cues; they are the foundation of our memories.

With respect to the coding of associations we have just discussed, I would now like to propose a model for how we store memories of our experiences in the hippocampus.[17] Let’s suppose we have a group of neurons that codes the concept of Luke Skywalker and another that codes the concept of Yoda. These characters are obviously related since they are part of the same film series, and such an association is represented by having some neurons that respond to both concepts.



Two groups of neurons representing the concepts of Luke Skywalker and Yoda. The association between these two characters, since they're part of the same film series, is encoded by neurons responding to both.

Technically (at least for now) we can't, as in *Total Recall*, manipulate the activity of neurons in the human hippocampus and see how this would affect the person's memory. However, we can do the opposite: We can manipulate memory and see if this manipulation affects the neurons' responses. More specifically, a few years ago we did an experiment in which we artificially created associations while recording the activity of these neurons. Let's suppose, for example, that we have a neuron that responds to Jennifer Aniston and not to the Eiffel Tower, and that we create a figure like the one below, where both concepts are combined, a type of selfie representing the (fictitious) memory of seeing Jennifer Aniston at the Eiffel Tower.



A neuron in the hippocampus initially responding only to Jennifer Aniston begins to respond to the Eiffel Tower when the association of these two concepts is generated.

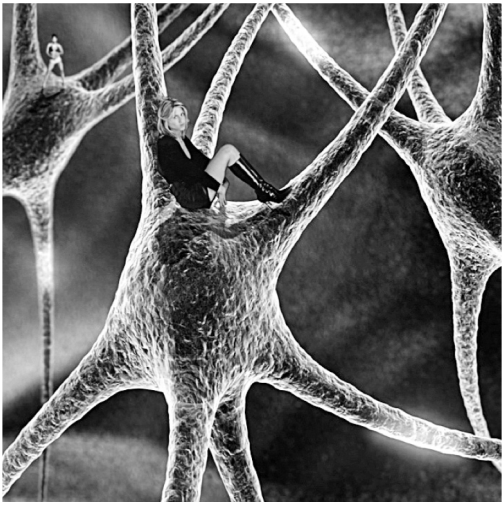
When the subjects learned this association (among some others they had to remember), we observed that the neuron that initially responded only to Jennifer Aniston also began to respond to the Eiffel Tower (without Aniston in the photo).[18] But the most interesting result was the speed with which these neurons changed their responses; it was at the exact time when they learned the association and sometimes after showing the "selfie" only once. This observation reinforces the idea of how these neurons are involved in memory formation: In fact, we don't live our experiences over and over again to be able to remember them, and we can remember a prominent event which only happens once, and perhaps in just a few seconds, for the rest of our lives. Although it is true that memories are constantly consolidated—or rather, reconsolidated—we need neurons, like the concept cells in the hippocampus, to quickly code these associations as soon as they happen.

These results have led me to conclude (I'm using the first person to clarify that this is my personal interpretation and not something that is universally accepted) that concept cells code associations, because the foundation of memory is precisely the coding of associations between concepts. My memory of that day at the beach with my brother is based on associating the involved concepts—brother, beach, Brazil. Although memory clearly involves many other aspects as well. I remember features of my brother's face, the sound of the waves, and the smell of seaweed on the coast. These memories are much more than abstractions or associations between concepts. But the representation of details happens in the cerebral cortex, particularly in the areas involved in the processing of information with the different senses (the details of a face are in the visual cortex, sound is in the auditory cortex, and so forth). The coding of details in the cortex is linked to the representation of concepts in the hippocampus, which allows us to relate different sensory impressions, for example, the smell, texture, and color of seaweed.

We have a conceptual representation in the hippocampus, a representation of "tags" with which it is easy to generate new associations.[19] If instead of concept cells we had a representation of details in the hippocampus, for each new association, for example, my brother and the Brazilian beach, we would need to establish connections among the details of these concepts, but without mixing them up with others, something that would be much more complicated since my brother resembles other people I know, and the Brazilian beach is similar to some Mediterranean beaches.

In general, arguing that the memory of our experiences is based on associations between concepts comes across as an exorbitant oversimplification. But this is probably due to a mistaken idea we have about memory, that is, believing that we remember much more than we really can. Beyond the representations stored in the cortex, many details we think we remember are perhaps nothing more than unconscious inferences we make using common sense. Perhaps I don't actually remember seeing my brother sweat under the February sun in Brazil as he sprawled on the sand like a lizard, but I infer that it probably was like that and this inference, which makes a lot of sense, becomes part of my memory and will enrich it every time I evoke it in the future.

Now, equipped with all the information we have discussed so far, let's get back to Akakhievitch's experiment. What would happen if we destroyed every neuron that coded Jennifer Aniston in the hippocampus? Would we completely forget the actress? This brings to mind the 2004 movie *Eternal Sunshine of the Spotless Mind*, where through a similar procedure, a couple decides to have each other completely erased from their memories after a dramatic fight.



What would happen if we eliminated all neurons coding the concept of Jennifer Aniston in a person's hippocampus?*

Based on the evidence provided by patient H.M., whom we introduced in the opening chapter, we know that the hippocampus is vital for memory; without it we couldn't form new memories. However, H.M. remembered things from his past: he knew who his mother was, his friends, the surgeon who operated on him. Similarly, returning to the example of the summer holiday with my brother, the memory of my brother and of Brazil—meaning, knowing who my brother is, where Brazil is, and so forth—resides in the cerebral cortex; what the hippocampus does is to quickly associate these concepts to later be able to remember the experience of that day at the beach. If I eliminated the neurons that code for my brother in my hippocampus, I would forget the details of that vacation and other associations that constitute memories of the experiences I have had with my brother, but I would still remember who he is.*

Selectively erasing the memory of my brother, or Jennifer Aniston, or any other person—meaning that after such a surgery we wouldn't know who they are—would be impossible because, as far as we know, there are no concept cells in the cortex selectively representing people, objects, or other specific concepts.** In the cortex, the representation is distributed—a millionfold democracy—and concepts are mainly coded by the same groups of neurons that fire one way or another to differentiate them. In other words, the cortex does not have a group of neurons that specifically fires to "Jennifer Aniston" and another group that responds to Maradona; there are no specific activations that we can eliminate to erase the memory of one person or the other. However, we could, in principle, manipulate the representation of each concept in the hippocampus and hence alter the associations underlying the memory of lived experiences.

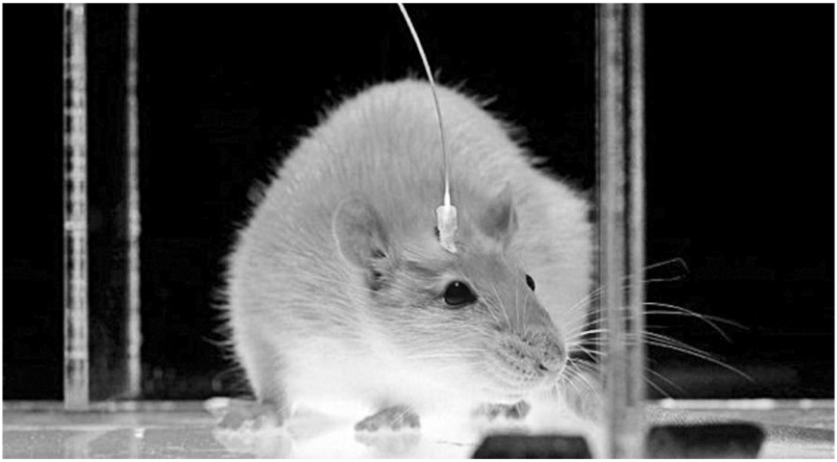
I'm rolling up my sleeves at my keyboard because I'm about to describe one of the most spectacular results in recent years in neuroscience. In an article published in 1979, Francis Crick (whom we've previously discussed) observed that practically any behavior we study involves a complex interaction of different areas in the brain and that, consequently, it would be invaluable to have a method that could selectively alter the activity of specific groups of neurons without modifying the rest of them. The underlying idea is to go from studying *correlation* to studying *causation*. For example, the activity of neurons in a certain area could correlate with a particular behavior but not be involved in its generation, not being its cause. Let's discuss this with an example.

A person's average income correlates with the size of their foot. It sounds ridiculous, but it is an undeniable truth. A person who wears EU size 43 tends to earn more than one who wears size 39; on average, the larger the foot, the greater the salary. The reason for this bizarre correlation is that men tend to earn more than women and usually have larger feet. The relationship with shoe size is a by-product, though it is real; we could randomly choose millions of people and obtain an incredibly high correlation. But if we wanted to prove that, in effect, the foot's size is what determines salary, we could try to find a way to make a group of people's feet grow (I wouldn't know how, but let's imagine it is possible) and see if this affected their income at the end of the month. This causality study, where we modify a variable and see how it affects another, would produce much stronger evidence about the relationship between the two effects compared to a simple correlation.

So, going back to Crick's claim, let's suppose we have neurons somewhere in the brain whose activity correlates with a specific behavior, for example, reading. However, it is possible that these neurons don't respond to reading itself but instead to the greater degree of attention required by reading. In other words, if a person did any task that implied paying attention, we would see the same thing. Ideally, we would like to perform a causality study and temporarily deactivate, like flipping a switch, the activity of these neurons and see if it is true that when we turn them off we can't read, and when we turn them on we can read again. Crick would later speculate that a potential method to control the firing of neurons could involve activations through light, since light can be turned on and off with great temporary precision. Decades later, it was demonstrated that he wasn't so far off the mark.

Certain algae have light receptors that allow them, using small moving filaments, to swim in the direction of light to accomplish photosynthesis like other plants. They do this via proteins that act as photovoltaic cells, which convert light into minuscule electric currents that guide the algae's movement. Setting aside endless technical details, neuroscience's great revolution in the last few years was to implant these proteins in the neurons of different animals, activating or inhibiting their activity when lighting them up.* To do this, the complication lies, on the one hand, in effectively introducing these proteins into the neurons and, on the other, in lighting them up to manipulate their activation. The first part was achieved by using advances in genetic engineering, in essence "infecting" the neurons with a virus that carried in its DNA the protein that reacted to light in the algae, and the second part was achieved through small fiber optic cables implanted in the brain.[20]

We have talked about Wilder Penfield and how he managed to evoke diverse reactions by electrically stimulating different areas of the brain. But the stimulation through light, what we call *optogenetics*, has many advantages because specific groups of neurons can be stimulated (without altering other neurons, which is hard to do with electric stimulation because, unless we introduce electrodes that touch a single neuron, every neuron around the electrode is indiscriminately stimulated), and the control over the activation and deactivation of neurons is much more precise. A little more than ten years ago, and almost thirty years after he proposed controlling the activity of neurons by just illuminating them —"turning them on and off" like a light switch—Francis Crick's dream came true.



Optogenetics: By lighting the inside of the brain using a fiber optic cable, genetically manipulated neurons can be activated or inhibited by illuminating them.*

In effect, initial studies showed that through optogenetics it was possible to change neural activity,[21] but the big question was if those manipulations could modify the animals' behavior, showing a causal role and not just a correlation between the activation of neurons and different functions of the brain.

The first results took a couple of years, but they were astonishing. A group of neurons in the mouse hypothalamus (which had been implicated in the regulation of periods of sleep and wakefulness and also in narcolepsy) was stimulated using optogenetics, causing sleeping mice to wake up immediately.[22] In another case, optogenetics was used to stimulate neurons in the motor cortex of mice, making them move their whiskers and stimulating a motor area related to locomotion; another study showed how they began to run like crazy each time the neurons in that area were lit.* But in the same way specific movements or periods of sleep and wakefulness are manipulated, could we also alter memories?

In patients with whom we performed recordings in the hippocampus, we saw that while making new memories, the activity of the neurons coding the involved concepts changed. The big question is if we can do the opposite: manipulate neural activity in the hippocampus using optogenetics to create new memories. This is very different from creating false memories through suggestion and reconsolidation processes. Now we are proposing something much more spectacular; we are talking about implanting memories through direct stimulation of the brain.

Japanese scientist Susumu Tonegawa received the Nobel Prize in 1987 for discovering genetic mechanisms that produce a variety of antibodies. Shortly after receiving the accolade, he radically changed his field of research from immunology to neuroscience. Since then, his lab has been reporting a series of amazing discoveries, particularly using transgenic mice—mice with genetic mutations that produce alterations in the activity of neurons and the connections among them—to clarify the role of the different areas within the hippocampus in memory formation.

The most astonishing results happened when Tonegawa and his team used optogenetics. In an initial study, they created conditioned responses in mice—a sound was followed by an uncomfortable electric shock—which prompted the mice to remain completely still when they heard the sound, anticipating the shock. But then they carried out the learning of the conditioned response using a very ingenious pharmacological/genetic manipulation. The mice received a compound blocking the incorporation of the virus with the photosensitive protein from algae, and the researchers then stopped giving them this compound while the mice learned the conditioned response. By doing so, they only infected the neurons related to this response—those that were active when the viral infection occurred—and therefore, the optogenetic stimulation produced when lighting the brain (specifically, in an area of the hippocampus called *dentate gyrus*) only activated these specific neurons.

Spatial context is crucial for mice. Knowing how to behave in one environment or another is a question of life or death; they survive by recognizing if they can move freely in a safe den or if they have to tread carefully due to the threat of a cat or other predator. Tonegawa's team managed to generate the conditioned response in a given environment and then placed the mice in a different environment, where they had never received an electric shock. The mice acted normally in this new "safe" environment, but when researchers "switched the light on" and activated the neurons involved in coding the conditioned response created in the first environment, the mice remained immobile, as if expecting a shock. In other words, when activating specific neurons, the researchers were able to evoke a memory and the behavior associated with it.[23]

The subsequent study was even more amazing. First, they used the same type of pharmacological/genetic manipulation to infect the active neurons with the virus containing the photosensitive protein while the mice explored a certain environment. Then they took the mice to another environment where, like earlier, they created a conditioned response (a sound followed by an electric shock), but while the mice learned this response, they turned on the light beam to activate the neurons representing the first environment, artificially creating an association between the first neutral environment and the appearance of the electric shock. Finally, they put the mice back in the first environment, where absolutely nothing had happened, and observed that they behaved as if they were in the environment where they had learned the conditioned response, showing the same type of freezing behavior they displayed when anticipating the shock. In other words, by activating specific groups of neurons, the researchers managed to generate a false memory: associating a neutral and safe environment with electric shocks that happened somewhere else.[24]

Let's go back to *Total Recall* and ponder how we could eventually turn Hauser into Quaid. To do so, we would have to erase the memories that connect him to Mars and his role as an agent for a government corporation that dominates the Red Planet, and also implant false memories, like being a factory worker and being married. In other words, we would have to erase some associations and implant others using, for example, optogenetics. What's interesting is that the implantation of false memories would basically be achieved by generating specific associations; so, in principle, it wouldn't be necessary to implant memories in detail since the brain takes care of that naturally, generating a construction around a few associations between concepts and enriching the memory with each evocation and reconsolidation, as shown by Loftus when implanting the false memory of the event at the shopping mall. It still sounds like sci-fi, although it is something that, in principle, is possible and likely not too far-fetched.

But can we use optogenetics in humans? Can we willingly manipulate the activity of specific groups of neurons and the behaviors associated with them, as shown in mice?

Optogenetics implies, on the one hand, the use of a virus to make neurons photosensitive and, on the other, the implantation of tiny fiber optics inside the skull to light them up. Apparently, this is far from being ethically acceptable, but there are many cases where such interventions would make sense. Recent studies are starting to show how manipulations through optogenetics offer very promising therapeutic alternatives to treat, for now with animal models, epilepsy, schizophrenia, anxiety disorders, addictions, depression, and Parkinson's, among other pathologies.[25] Therefore, it wouldn't be surprising if the advances demonstrated with animals were soon applied to humans.

Remember our friend Akakhievitch? Let's now imagine that his grandson, Drogori, after reading about the advances and great challenges in neuroscience in recent years, decides to defend his grandfather's name, so tarnished after Lettvin's story. Almost no one takes the idea of "grandmother neurons" seriously, because not even concept cells in the hippocampus are close to doing what his grandfather Akakhi stated. But Drogori conceives an even more brilliant and diabolical idea. Drogori already knows about Loftus and Tonegawa's experiments; he knows about the possibility of erasing and implanting memories, but what truly fascinates him is understanding what distinguishes humans from other animals.

The advances in artificial intelligence in recent years have made us reconsider what defines our intelligence and wonder if one day we will create silicon beings that may surpass us. But computers still lack general intelligence; they can't make abstractions, extract sense from things, or make inferences and analogies. Computers reliably store everything without selecting or elaborating on essential information. However, it wouldn't be surprising if in some years some IT genius could find a way to make computers understand the content of what they store, replicating the processes done by the human brain. Therefore, what truly intrigues Drogori is the difference between humans and other animals, especially higher primates. After all, their brains are similar, but monkeys lack concept cells; or to be bolder, the existence of concept cells is probably the key difference between the monkey and the human brain.*

Fascinated by the *Planet of the Apes* story, where a chimpanzee, Caesar, begins to think and speak like a human, Drogori comes up with a singular experiment to implant concept cells in monkeys. He knows that the lack of these neurons in animals is likely due to the fact that they don't have language, which is what reinforces the representations of concepts, but Drogori has a shortcut in mind to achieve this, using optogenetics and experiments inspired by those done by Tonegawa. His idea is to selectively activate a group of specific hippocampal neurons in monkeys while showing them pictures of, for example, one of the vets who takes care of them. The activation will generate an engram, a group of neurons that will respond to the vet. But the key lies in not showing just one, but a variety of different pictures of the vet—face forward, in profile, wearing lab scrubs, in casual clothes, in a suit, and so on—so as to create a representation of the concept rather than a specific picture. In other words, his idea is to induce neurons that respond to the vet regardless of how he looks. Furthermore, Drogori plans to not only use images but also auditory stimuli, like the sound of the vet's voice, and even olfactory stimuli, with his particular smell, in order to generate a multimodal representation combining the different senses.

Drogori's results would revive his grandfather's old idea—generating neurons that respond to a specific person—but his objective is far more ambitious. Just as he could create an engram representing the vet, he could also create others representing the other monkeys in the habitat, the people in touch with the monkeys, different types of food, and other concepts present in their daily lives. Basically, Drogori would be creating an explicit neural representation of concepts in the hippocampus, which the monkey could eventually begin using to form new associations and episodic memories. In other words, he would generate an abstract representation of the things like the one we use in our own thought processes. The big question Drogori asks himself, then, is if by creating these representations—something that is in principle doable, since it is an experiment similar to those performed with mice—he could cause the monkey to think and even develop human intelligence. In a book about science and science fiction, I allow myself this fictitious and somewhat reckless (though possible) argument to close this chapter.

---

* The movie was released in 1990 (and re-released in 2012 with some plot variations) and is based on a story, once again, by Philip K. Dick, titled, "We Can Remember It for You Wholesale," first published in 1966.

* Borges's "La memoria de Shakespeare" was originally published in 1983 in a story collection by the same title. Norman Thomas di Giovanni's English translations are considered paramount, although sadly out of print.

* Héctor Maldonado, a pioneer in neuroscience and the study of memory in Argentina, argues that it is no coincidence that Mnemosyne is the mother of the muses, since artistic inspiration emerges from memory (*La memoria animal*, Buenos Aires: Eudeba, 2008).

* The left image is of a wax tablet from Ancient Egypt, exhibited in the Metropolitan Museum of Art; the reproduction of birds in a cage by Canadian ornithologist and artist Allan Brooks (right) appeared in *National Geographic*.

* These findings had significant repercussions because they highlighted the subjectivity of eyewitnesses during oral trials and how a testimony can be manipulated according to the way a question is asked. For more information, see chapter four of *The Forgetting Machine* (cited earlier) and Elizabeth Loftus, "Our Changeable Memories: Legal and Practical Implications," *Nature Reviews Neuroscience* 4, no. 3 (2003): 231–234.

Moreover, the malleability of our memories during reconsolidation warns us of the power of political propaganda and the repetition and subjective interpretation of news in the media. This is closely linked to the discussion we had at the end of chapter six about blind choice and how easy it is to manipulate decisions.

* William James, *The Principles of Psychology*, vol. 1 (New York: Henry Holt and Company, 1890), chapter 6. Similar ideas were also proposed by Santiago Ramón y Cajal, who spoke of psychic neurons in the cortex, responsible for complex mental functions, and Sigmund Freud, who posited the existence of Psy and Phi neurons, in charge of perception and memory, respectively.

** Sir Charles Sherrington, *Man on His Nature* (Cambridge: Cambridge University Press, 1940). Sometime later, British neuroscientist Horace Barlow argued that the brain doesn't make decisions based on a single *pontifical cell* or a *millionfold democracy*, it rather uses a *college of cardinals*, meaning the brain does not use unique neurons or distributed networks but a representation with relatively few neurons (but not just one). This is what in neuroscience is known as a *sparse* representation. For more information, see Horace Barlow, "Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology?" *Perception* 1, no. 4 (1972): 371–394.

* Aristotle's *On Memory and Reminding Oneself* was written around 350 BC.

* This image was created by my colleague Dave Barton to illustrate a talk I once gave about the Jennifer Aniston neuron at a Café Scientifique event in England.

* The distinction isn't quite so cut and dried, since some associations, at first *episodic* (established in the hippocampus), can also become part of the *semantic memory* stored in the cerebral cortex. For example, I can remember the event of seeing my brother with his partner for the first time, which is an episodic memory, but after many other meetings, I will generate a semantic, factual memory of knowing who my brother's partner is, without necessarily remembering that first meeting (although in this case, I clearly remember different associations of that memorable night).

The discussion regarding what memories and associations are transferred to the cerebral cortex during the consolidation process and which ones remain stored in the hippocampus has led to heated debates in neuroscience, resulting in two opposing models: The Standard Consolidation Model claims that both episodic and semantic memories are consolidated in the cortex and that the hippocampus only has one temporary function in memory formation, while Multiple Memory Trace Theory argues that episodic memories remain permanently coded in the hippocampus.

For more details, see Morris Moscovitch et al., “Functional Neuroanatomy of Remote Episodic, Semantic and Spatial Memory: A Unified Account Based on Multiple Trace Theory,” *Journal of Anatomy* 207, no. 1 (2005): 35–66.

** To be clear, I’m talking about selectively erasing a concept while leaving others intact. Experiments in rats have proven that memory can also be erased but without such specificity, inhibiting a protein known as *protein kinase C, zeta* or “PKMζ,” which is fundamental in maintaining connectivity among neurons (representing memories). For more information, see Eva Pastalkova et al., “Storage of Spatial Information by the Maintenance Mechanism of LTP,” *Science* 313, no. 5790 (2006): 1141–1144; and Reut Shema, Todd Charlton Sacktor, and Yadin Dudai, “Rapid Erasure of Long-Term Memory Associations in the Cortex by an Inhibitor of PKMζ,” *Science* 317, no. 5840 (2007): 951–953.

It would be a different case entirely if future experiments indeed show that concept cells are also present in the *neocortex*. We currently don’t know, as the possibility of performing single neuron recordings in the human neocortex is quite limited due to clinical and technical reasons.

* The activation of neurons is accomplished using a protein called *channelrhodopsin*, while another, called *halorhodopsin*, is used for inhibiting them. Currently, proteins that have different effects on the activity of neurons have been developed, and they are activated with lights of different colors (blue, yellow, green). It sounds incredible, but after carrying out the genetic process of introducing these proteins into the neurons, different types of manipulations in their activity can be induced according to the color of the light used.

* Image credit: John Carnett, Viviana Gradinaru, Murtaza Mogri, and Karl Desisseroth of Stanford University.

* The experiment that made mice move their whiskers was conducted by Alexander M. Aravanis et al., “An Optical Neural Interface: In Vivo Control of Rodent Motor Cortex with Integrated Fiberoptic and Optogenetic Technology,” *Journal of Neural Engineering* 4, no. 3 (2007): S143–S156. For a demonstration of how the stimulation in a motor area made the mouse run like crazy, see “Fiberoptic Control of Locomotion in ChR2 Mouse.” YouTube video, 00:28. Posted by “eekolife,” November 18, 2008. www.youtube.com/watch?v=88TVQZUfYGw.

* In chapter three, we discussed the similarity among place cells found in the hippocampus of mice, and concept cells in the human hippocampus. However, place cells respond differently if the context is changed and therefore have a limited capacity for abstraction, contrary to concept cells, which respond the same way in different experiments and contexts. For more details, see a short article of mine, “Neural Representations Across Species,” *Science* 363, no. 6434 (2019): 1388–1389.

In monkeys, the closest to these neurons are the ones that respond to faces in face patches, mentioned in chapter five. However, it has been recently shown that these neurons don’t respond to specific people but to visual details of different faces. For more information, see Le Chang and Doris Y. Tsao, “The Code for Facial Identity in the Primate Brain”; and my article “How Do We Recognize a Face?” (both cited in chapter five).

Recently a group of researchers performed recordings in the monkey hippocampus, replicating the experiment with which we obtained concept cells in humans, but they didn’t find this type of neuron. For more information, see Julia Sliwa et al., “Independent Neuronal Representation of Facial and Vocal Identity in the Monkey Hippocampus and Inferotemporal Cortex,” *Cerebral Cortex* 26, no. 3 (2014/2016): 950–966.

CHAPTER 10

OPEN YOUR EYES (VANILLA SKY)

Immortality

Candela
Insua

Generations of Argentines grew up reading stories of extraordinary characters in magazines published by Columba: *El Tony, Intervalo, Fantasía, D'artagnan, Nippur Magnum.** In May 1969, *D'artagnan* released "Gilgamesh el inmortal" ("Gilgamesh, the Immortal"), a comic by illustrator and scriptwriter Lucho Olivera, which initially was due to have just one episode but, due to popular demand, turned into a series with more than 140 installments.

The story is based on a beautiful Sumerian legend in which Gilgamesh, king of Uruk more than two thousand years before Christ, sets out on a journey across deserts, seas, and mountains to find the secret to eternal life. After a long voyage, he finds Utnapishtim, the Distant One, the only man who was granted immortality by the gods, and he reveals to the wandering king that the secret to eternal life is in a Lycium plant found at the bottom of the sea. Gilgamesh ties rocks to his feet to reach the bottom of the waters, and once he finds the plant, he unties the rocks and gets back to his ship with the precious elixir. However, during his return trip, a serpent steals the Lycium plant, stripping him of the gift of immortality (although, according to the legend, Gilgamesh ruled Uruk for 126 years).

Our lives go by too fast. We live a few decades, and sooner or later we face the dreadful phantom of death. Yet science's extraordinary advances brazenly defy the inevitability of our extinction and, like Gilgamesh in ancient Uruk, we search for the Lycium flower to perpetuate ourselves till the end of time. It is the strongest impulse of every living being, our *conatus*, the survival instinct rooted within each of us that makes us fear death and cling to life.

The notion of Cartesian dualism, advocated by Western religions, offers great comfort: With death it is the body that perishes, but the mind, the soul, that which defines our being, survives material extinction and makes us immortal. Yet, as we discussed in chapter one, Cartesian dualism is hard to sustain from a philosophical point of view. Although it is worth noting that science, reason, cannot demonstrate the existence or nonexistence of a god or life beyond death; that's a question of faith. But let's set aside religious arguments and analyze concrete possibilities that science offers us to transcend death.

Currently, most scientists embrace the idea of materialism: The mind does not exist independent from the body; rather, they are one and the same. Therefore, there's a material substrate—neural activity—that determines memory and identity. And this is precisely the fact that opens up the possibility of achieving the apparently impossible, to defeat death in the great chess match, as famously depicted by Ingmar Bergman in *The Seventh Seal*.

Carried along by the most formidable scientific crusade of all times, let's then imagine that we can maintain or, more specifically, cryopreserve our body when it perishes; in other words, preserve it at a very low temperature to avoid the decomposition of organs, so as to eventually resuscitate it when scientific advances allow, in dozens or maybe hundreds of years. After all, who can assure us this won't be possible in the future? Until a few years ago, it also seemed impossible to create cyborgs or computers that could defeat chess champions and, going further back, the idea of humans flying to the moon was a pipe dream.

The cryonics idea has been used in countless sci-fi movies,* among them *Open Your Eyes*, by Alejandro Amenábar, where a person is preserved in a continuous virtual sleep waiting to be resuscitated.* Up to this point, there's nothing new—Amenábar's movie is one of many using the same story line, but with a major difference: *Open Your Eyes* outlines what it would be like to remain asleep for decades, giving way to an exquisitely complex plot.

Scene from Ingmar Bergman's *Det sjunde inseglet* (*The Seventh Seal*), in which a medieval knight challenges Death to a chess game in order to prolong his life. (© AB Svensk Filmindustri (1957), Photo: Louis Huch, Stills Archive: The Swedish Film Institute.)

Earlier we wondered if perhaps some nefarious mastermind, Descartes's famous evil genius, could create the illusion of a nonexistent reality by stimulating our senses. We recall a similar idea explored in *The Matrix*, where two parallel realities are combined: the one of the Matrix, with humans in incubators experiencing a nonexistent reality, and the one faced by those who have freed themselves from that false reality and fight the machines that generate it.

This chapter's opening illustration is inspired by the opening sequence in *Open Your*

Eyes, where the main character, César, finds himself walking down Madrid's Gran Vía, completely alone. The character's solitude, in a city that is usually crammed with people, decidedly stands out and evokes Dan Dennett's argument that the idea of *The Matrix*, or Descartes's demon, is actually impossible because we cannot replicate the infinitely complex feedback we have while interacting with other people and the outside world. The Gran Vía scene is, in the end, nothing more than a dream, and once César wakes up, the scenario repeats itself in the daily chaos. The crosscutting is perhaps the most captivating aspect of the movie, visually narrating the story's masterful fusion of reality and imagination, wakefulness and sleep; and how memories and even people's identities are jumbled for someone experiencing eternal sleep. The complexity of the outside world and our interactions with it shows us then that external reality exists, while the coherence of the facts that happen to us reinforce that we are not dreaming; regardless of how elaborate the dream we are immersed in may be, the illusion cannot be indefinitely sustained. At some point the dream narrative would start to fail, as it does in *Open Your Eyes*.

The plot refers to a cryonics company in Arizona called Life Extension. Oddly enough, such a company exists. Its name is Alcor Life Extension Foundation and, as in the movie, it offers the service of preserving the body in liquid nitrogen after death, allowing for the possibility of regaining the person's consciousness when technology grants it in the future.*

The process is much more complex than merely decreasing the temperature to freeze the body. This is how it works: Once the person has been declared dead,* the corpse is placed in an ice bath to quickly reduce the body's temperature, and respiratory assistance is applied to avoid damage caused by the lack of blood flow, particularly in the brain; then a blood thinner is injected into the body's bloodstream; and finally the blood in the circulatory system is replaced with a chemical compound that, when decreased to lower temperatures, becomes a viscous liquid until it reaches a glassy state. This process, called *vitrification*, is done to avoid the crystallization of frozen water, which can tear the body's tissues. Finally, the body is preserved at −196°C (the temperature of the liquid nitrogen used in this process, the equivalent of −320.8°F), avoiding any type of chemical reactions and guaranteeing its preservation for centuries.[1]

ALCOR
LIFE EXTENSION FOUNDATION SINCE 1972
www.alcor.org

A cryonics tank for up to four bodies from the Alcor company in Arizona.**

It should be noted that the cryopreserved brain can't actually dream, like in *Open Your Eyes*, because the neurons remain frozen and inactive. We saw that sleep is much more than a resting period, and that it has a very active and essential role in learning and, especially, in memory consolidation. During sleep, neurons don't just rest. On the contrary, they are very active, producing all the fantastic stories we barely remember when we wake up. Therefore, instead of cryopreserving the brain, we should keep it alive (and active) to allow the person to live a perpetual dream.* With cryonics, we wouldn't be able to dream because we would be clinically dead, with absolutely no brain activity.

So, let's set aside the possibility of pleasant sleep until being resuscitated and discuss an even more fundamental question: Does it makes sense to revive a frozen person? At first, cryonics seems far-fetched and is considered by many as pseudoscience. However, it is based on some scientific evidence and on widely used transplant techniques. In fact, different organs are routinely frozen and preserved at low temperatures, so it is worth asking ourselves if we couldn't do the same with a person's brain and identity.

In the 1950s, it was demonstrated that it was possible to revive rats that had been taken down to 0°C and did not breathe or have a heart rate for more than an hour.[2] However, the freezing process at liquid nitrogen's temperature could harm the brain's structure. The problem is that the cryogenic compounds are highly toxic and, additionally, when using them on complex organisms, the possibility of controlling their diffusion into different tissues diminishes. Furthermore, experience in organ preservation for transplantation indicates that the process varies according to each organ to be preserved, requiring different freezing speeds and different concentrations of cryogenic liquids; therefore, the vitrification process of a body is far from trivial.

However, through the development of a new cryonics technique, researchers in California recently preserved a rabbit's brain at −135°C and when defrosting it, analyzing brain slices under a microscope, they showed that neurons and the connections among them had remained unharmed.[3] In spite of being an encouraging result for cryonics advocates, there are still plenty of problems to resolve. Specifically, analyzing a rabbit's brain slices is very different from bringing the animal back to life and proving that it behaves as it did before being cryopreserved. Moreover, we have absolutely no idea how we could resuscitate an animal or a person after death (which is very different from freezing an animal for an hour). If we knew how to do it, we would already be implementing this at hospitals. We wouldn't have to freeze people and wait for scientific advances in the future to revive them; we would resuscitate them as soon as they die. (And if in the future we can revive a person who is brain dead, we would have to reconsider how we define *death*; in the capitalist society we live in, it wouldn't be too outlandish to think that we might end up considering someone to be dead when they no longer have enough money to pay for complex resuscitation treatments.)

Cryonics supporters, however, argue that although we don't know how to resuscitate a person, that doesn't mean we won't be able to do it in the future. After all, in the past, a person who had a heart attack was considered dead, and nowadays we can revive them through widely established procedures. But we must clarify that cryonics is still science fiction, just like the possibility of taking intergalactic trips using space-time warps. What contemporary science does have a handle on is the cryonics process itself, in other words, how to bring body tissue to extremely low temperatures without harming it. But this is far from something that's been proven and established. In fact, it wouldn't be surprising if future scientists aren't able to even consider how to revive our bodies due to irreparable damage to our tissues, especially the brain, that occurred during the freezing process.

Despite the onslaught of criticism, many people believe it is better to have a chance, regardless of how tiny it may be, than absolutely nothing, and nowadays there are already hundreds of people whose bodies (or only their heads) are preserved in tanks at low temperatures, and thousands more who have registered to receive the treatment as soon as they pass away.* It is a big bet. But who would dare to predict scientific advances? Beyond the remote possibility of resuscitating a cryopreserved body, there are other basic factors that could interrupt the process. Nothing guarantees that the bodies will be preserved for centuries because they could be destroyed by an earthquake, nuclear war, or simply because the company offering these services goes bankrupt.

Let's now raise an even more essential question, picking up the *RoboCop* identity discussion: Would these people be the same once they were resuscitated? Or is cryonics an extremely expensive and complex procedure to create a brand new person in the future, rather than bringing back the original one?

We saw that the notion of identity is linked to the dilemma of change: Things change, but they continue to be the same. We ourselves change and still remain the same person. But what is it that remains the same while we are alive and is lost with death?

At Christmastime in Sweden, a seven-year-old girl unexpectedly left home to search for Santa Claus. With an outside temperature of −12°C (10.4°F), she fell down a cliff and lost consciousness, submerged in freezing waters. She was rescued more than an hour later, but her situation was severe: she was in a coma, her body temperature was 13°C

(55.4°F), and she had no vital signs (she wasn't breathing, had no heart activity, and her pupils were dilated and unresponsive to light). Astonishingly, after hours trying different treatments to revive her, she began to breathe intermittently, her heart began to beat again, and the following day she was out of her coma.* This is one of many resuscitation cases of people with hypothermia due to being submerged in freezing waters, snow confinement (for example, after an avalanche), or people who live on the streets and are unable to withstand an intensely cold night.

The Swedish girl, as well as others who recovered their vital signs several hours later, are the same people as always. The loss of consciousness due to the freezing of their bodies didn't make them come back to life with another identity. However, it should be noted that in such cases the body temperature doesn't fall below 10°C, far from the −196°C used on people in Alcor's cryonics. But beyond technical difficulties related to vitrification, let's for now accept that preserving a body at 10°C or −196°C shouldn't change anything regarding the person's identity.

But what about a mountaineer, such as George Mallory, who disappeared while climbing Everest in 1924, and whose body was found seventy-five years later?* Even if we defrosted him and performed CPR, we wouldn't be able to revive him (it seems utopic, but this is a relatively easy task compared to resuscitating a cryopreserved body at −196°C). But let's suppose that in a couple of centuries, if the resuscitation of frozen bodies were possible, we find the body (until now still lost) of Andrew Irvine, who disappeared with Mallory during the 1924 expedition. Applying the latest scientific advances, let's imagine that future scientists recover Irvine's heart rate, breathing, and reflexes, and after a while, they manage to resuscitate him. Would he still be the same person? Would he feel as if he had just woken up from a prolonged coma? And if we argue that he is not the same, what's the difference between him and the Swedish girl? Why are we still the same after one hour and not after several years of being frozen?

Perhaps a fundamental difference between hypothermia cases and that of frozen mountain climbers is that patients with hypothermia, in spite of not showing vital signs (breathing, heart activity, and associated responses), likely have had brain activity before being resuscitated.** During our entire lifetime, our brain never stops; the neurons never cease to be active, not even in the deepest sleep. This is not the case with the frozen mountain climbers in the Himalayas or those in Alcor's tanks awaiting resuscitation. In their cases, the brain's continuity has been interrupted, and it is not clear at all if they would still have their identities if they're eventually resuscitated.

Andrew Irvine (*left*) and George Mallory (*right*), two mountain climbers who disappeared in 1924 while trying to become the first to summit Everest.

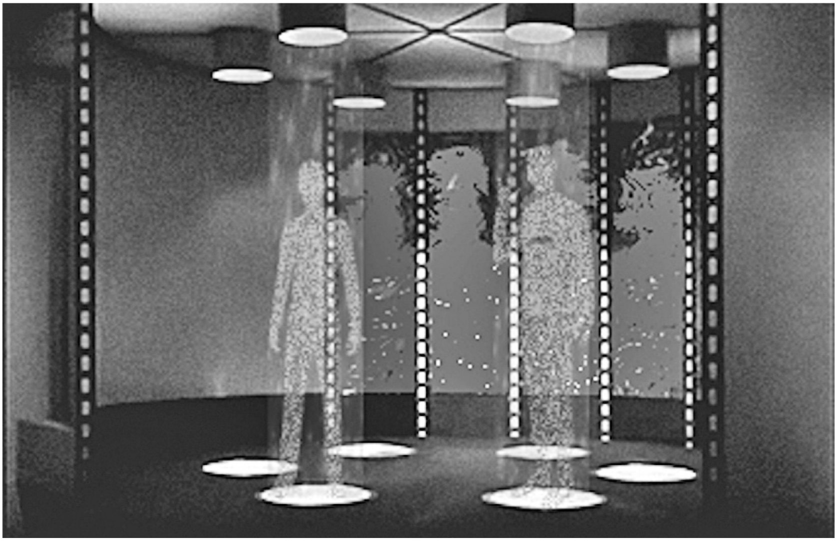
Let's continue discussing this problem, but analyzing another variant of perpetuating ourselves provided by sci-fi. To do so, let's briefly revisit Theseus's paradox, described in chapter seven, but using a very interesting variation proposed by Thomas Hobbes (1588–1679). Let's imagine that as Theseus's ship is being restored, someone commits to gathering the discarded planks and rebuilds the vessel at a distant port, using the original pieces. Which is Theseus's ship then? If we put ourselves in the shoes of the people living in Ancient Greece, we would probably consider his ship to be the one that leaves every year for Delos Island. To them, the fact that someone gathered the original planks and rebuilt the ship somewhere else wouldn't change anything; the second boat would be nothing more than a curiosity.

This is the basis of *functionalism*, which contends that what matters is the function of things rather than what they are made up of. Applied to the brain, functionalism says that mental states are determined by their function and not by the material substrates that generate them.

So, let's imagine that one day, rather than cryopreserving our bodies, we are eventually able to reproduce our brains.* In other words, instead of keeping our brains inert for centuries, let's consider the possibility of copying each of our neurons and the connections among them. This would be, by the way, a colossal job, since we have around 100 billion neurons and each neuron is connected to an additional ten thousand.** The number of possible connections goes way beyond what we could deal with nowadays, but let's imagine that in the near future science could make it happen. To be more specific, let's imagine that we reproduce the brain's entire architecture in a supercomputer, that is, replacing neurons with transistors in tiny chips and connecting them in exactly the same way as in the original configuration. Following the principles of functionalism, the fact that cerebral activity is generated by carbon circuits (organic matter) or silicon (inert material used in computer chips) shouldn't change anything and, consequently, this supercomputer's configuration would store our memory and identity. But would it really be us?

The idea of synthetically reproducing a person—and the possibility of transcending death by doing this—has been vastly explored in science fiction.* Undoubtedly, one of the most popular phrases among this genre's fans is, "Beam me up, Scotty," Captain Kirk's repeated request to Montgomery Scott, the *Enterprise's* chief engineer, to be teleported in the TV series *Star Trek*. The principle of teleportation consists of disintegrating a person in their original location and reintegrating them almost instantly in a different place.** So far, so good. But the fifth episode of the show's first season brings up the following situation: Due to a glitch in the transporter, a copy of Captain Kirk is created with a completely different personality.[4] The replica behaves strangely, and we can immediately recognize the imposter. However, in an episode thirty years later, the problem becomes more interesting.[5] In this case, Commander Riker encounters a copy of himself on planet Nervala IV, who had been waiting for eight years to be rescued. This situation happens due to a glitch in the transporter, which during a previous visit to that planet brought Riker back to the spacecraft, but without disintegrating the one remaining on Nervala. Who is who? Both Rikers have equally valid arguments for being the original, one because he continued his routine life on the spacecraft and the other one alleging that his crew abandoned him. This episode's plot specifically plays with the fact that it is impossible to choose one over the other.

The arguments and dilemmas caused by the idea of teleportation are so interesting that they have attracted the attention of contemporary philosophers. One of them was Derek Parfit, who used teleportation to explore a dilemma regarding identity.[6] He first considered the following scenario: Let's imagine we enter a transporter and, when pressing a button, the machine scans the information of every one of the particles that constitute our body (including our brain) and then destroys it in the departure location to reconstitute it at the arrival destination, using the information transmitted by the transporter.* So, we are about to take our first "teleporter trip." Although we feel somewhat nervous, we want to see Mars at all costs; in the end, our curiosity trumps our fears and we press the button. Soon after, we appear on a Mars colony and we observe the surrounding reddish landscape in awe, happy that everything worked out. But did it work?



Teleportation. A person's body is disintegrated in an original location and instantly reintegrated at the destination site.

To make teleportation's problem more evident, Parfit proposes a second scenario: Years later, we repeat the experience. We return to the transporter, press the button, but this time nothing happens; we're still on Earth, waiting to be teleported. We leave the machine and let the technicians know that it is not working. However, they tell us that the teleportation did work and that we're already on Mars. A machine error delayed the disintegration at the original location, but we shouldn't worry because we will soon be disintegrated. With growing despair, we say we are still there, on Earth, that we weren't teleported. However, they show us images of Mars where we have just arrived. The person on Mars is exactly like us, behaves like us, and can answer every question about our past, even recalling memories that we never shared with anyone. This is a very different case from Quaid and Hauser in *Total Recall*, because now the "original" on Earth and the "replica" on Mars don't just have the same body but also the same memories. There's an exact copy of us on Mars, someone who will be just as loving with our family, who will try to finish the book we have been trying to write for years, but this brings no comfort because, in the end, we are still there, on Earth, and we don't want to die. We want to be ourselves, and not that other person on Mars, the one who will see our children grow and witness our future.

With this example, Parfit aims to demonstrate that teleportation is a death sentence. The original dies in the disintegration process, while the reintegrated person at the chosen destination wouldn't realize that they have just been created. In fact, they wouldn't notice anything strange and would believe themselves to be the same person who pressed the button on Earth because they have the same memories. In conclusion, if one day science allowed us to make an exact copy of our body (and our brain), that wouldn't mean we would perpetuate ourselves. We would produce a perfect replica of ourselves (which won't realize it is a copy), but we would cease to exist and would not witness what that copy will do the rest of their life. And just as we have our doubts about teleportation, it is also natural to wonder if we would maintain our identity through cryonics or a replica of our brain in a supercomputer.

Parfit continues arguing that the people who show up on the other end of the transporter did not experience our original memories but merely believe they did so because their brain (and the memories stored in it) is an exact copy of our own, and there's no way of establishing the difference. And here lies the conundrum: the replica has our memories, but it is not us. With this reasoning, Parfit concludes that memory cannot explain identity. On the other hand, we saw in chapter seven that we are also not the continuity of our body. What then determines our self?

The only way Parfit is able to exit this crossroads is by concluding that the self doesn't exist as such, that it is an illusion. The body obviously exists, as well as the brain and all the memories stored in it—the illusion is the person's sense of identity. To illustrate this, Parfit considers a club with members that gather regularly. After a few years, those meetings stop happening, and a while later some members decide to get together again like they did earlier. Is this the same club as always, or is it a new club? Parfit views this as an empty question, one that makes no sense: What matters is that the members got together and interacted among themselves again; whether we consider it to be the same club or a different one is irrelevant.

Parfit's view is not new. It is the *anatta*, or denial of the self, preached by Buddhism. When a king once asked a Buddhist monk his name, the monk replied that he was known as Nāgasena, but that Nāgasena was only a name because it didn't represent a person. According to Buddhism, actions and consequences exist, but not the person who is acting.[7] In Western philosophy, Hume also proposed the denial of the self:

> For my part, when I look inward at what I call myself, I always stumble on some particular perception of heat or cold, light or shade, love or hatred, pain or pleasure, or the like. I never catch myself without a perception, and never observe anything but the perception . . . [Someone] may, perhaps, perceive something simple and continued that he calls himself, though I am certain there is no such thing in me.
>
> —David Hume, *A Treatise of Human Nature* (Book I: part 4, sec. 6), 1739

For Hume, the self, the notion of identity, is nothing more than a bunch of perceptions happening one after another. Continuing this line of thought, Dan Dennett says that identity is an illusion created by our narrative: It is the language that leads us to talk (and talk to ourselves) and think "as if there were someone inside."[8] Moreover, we earlier saw that our body's sense of belonging is relative and occurs because of the coherence of our feelings: We believe the illusion that a rubber hand is part of our body, or we believe the opposite if an "alien" hand doesn't behave according to our predictions. According to this perspective, we are nothing more than sophisticated organisms, but so complex that we can reflect on our own behaviors and even be conscious of our existence, creating the illusion of a self. And if the self doesn't exist, wondering about death makes no sense because there was never someone who would die.

The denial of the self is much easier to accept if we abstract ourselves and analyze the problem in the third person, that is, if we stop thinking about what would happen to us and see it all from the outside, say from the perspective of a Martian who arrived on Earth and is trying to understand the philosophical problem that is the bane of our existence. Returning again to the world of science fiction, this is precisely how we see the army of clones in *Star Wars*. We don't wonder about the identity of each of them; it is much simpler: They are a group of clones, with the same body and brain, living independent lives. Likewise, the person who remains in the transporter on Earth and the one who arrives on Mars—like the Riker on the *Enterprise* and the other one who remains on Nervala IV—are identical copies who carry out their lives separately.

Star Wars' army of clones. By not putting ourselves in their shoes, we don't wonder about their lives, identity, or fear of death.

Nonetheless, the idea that the self doesn't exist is not an easy pill to swallow. It is a blow to what we consider philosophy's first truth—*I think, therefore I am*—proposed by Descartes. The argument is still valid, but it radically changes the conception of what exists, since it would no longer be the self. But the feeling of existing, of being a person, is too strong. In fact, it is what I believe I know with the greatest certainty. It is not based on any rational argument; the feeling that there's a self writing these lines is in my gut. And just like in a famous dialogue written by John Perry, where the character Gretchen Weirob is not satisfied by general arguments and starkly wonders about the continuity of her own life before dying,* we can't avoid analyzing the problem in first person. That's the perspective of movies such as *Moon* (2009) or *Oblivion* (2013), where we identify with the main character and his attachment to life—be it a person who supervises automated facilities to extract fuel on the moon, or a technician repairing drones that guard energy generators on the already abandoned planet Earth—only to find out he is just another clone. These movies skillfully remove us from the third-person position we take with *Star Wars*—except *Episode VII, The Force Awakens* (2015), where a clone also wonders about himself—and lead us to ask the same questions these clones wonder about, the same questions we have been asking ourselves for centuries.

To Parfit, asking about a person's continuity is like asking if a precious gold watch that we are getting repaired is the same after having been taken apart and put back together by the watchmaker; there's nothing to explain, the discussion is as irrelevant as asking how we should consider a group of people who have gotten together again. However, there's a fundamental difference. Neither the club nor the watch are aware of themselves. In these cases, there's no alternative beyond viewing the problem from the outside; we can't put ourselves in the club or the watch's place because they aren't conscious beings. But if we think about a person, there is something that does have existence, even if it is only "nominal" (as stated by Buddhism). In other words, even if we accept the idea that the self is an illusion—actually, I would rather say a *construction* of the brain, like the mind or free will—it makes sense to ask ourselves about what determines the feeling we have of our own existence, of being the same person throughout time.

The feeling of existing, of being conscious of ourselves, is similar to qualia, the feeling we have when seeing the color red or smelling jasmine. Once again, we return to the hard problem we discussed in previous chapters. We don't perceive electromagnetic waves of a certain frequency, or a specific chemical composition of molecules; *we feel* the color red and the smell of jasmine. Similarly, we don't tend to perceive a handful of feelings (although introspectively we may be aware of them) but rather our self. Going back to the recordings we have performed with epileptic patients, in some cases we find neurons that respond to pictures of the patients themselves, meaning the patients have a conceptual representation (and not of details or isolated feelings) of themselves as a whole.

In my view, with the teleportation example Parfit demonstrates that there doesn't necessarily have to be just one unique self; infinite copies of one person could coexist, and each copy would have a notion of self that will diverge with time, given each copy's different lived experiences. Once again, the paradox happens when we analyze the problem in the first person, taking the *Moon* or *Oblivion* clone's point of view. That's when we must ask ourselves: What makes me be right here and not where my clone is, if we are exactly the same? Should we then reconsider materialism and accept that there is something beyond matter that determines that we are in one place and not another? The answer is that I am both here and there, and not because I have some sort of connection with my clone, but because there isn't a unique self, and analyzing the problem from the outside, in the third person, we would realize that the clone is asking himself the same question, considering that I am a copy. On the other hand, the illusion of the identity's continuity, of being the same person throughout time, is in effect given by memory. This organism that I call "myself" remembers what happened to it in the past and can plan its future based on that. Through memory we create a story, our story. Therefore, replicating our memory wouldn't allow us to perpetuate, as Parfit showed us with his teleportation examples, but it is memory that creates a unified and coherent view of our existence. Memory (even though it is initially the same) is what makes each soldier in the army of clones feel like a person.

Let's add a twist to Parfit's thought experiment. We already saw that the transporter's glitch would make the self on Earth ask about the abominable creation that appeared on Mars, just like the self on Mars, the one that experienced teleportation, will ask about the "waste" that remained on Earth. But let's imagine that the transporter works well but doesn't take us anywhere. It simply destroys our body (and our brain) and rebuilds it a while later, in the exact same location. Let's imagine that we disappear and reappear after a split second. Will we be the same?

Let's now imagine that the reconstruction happens after several years. Oddly enough, this would be similar to resuscitating someone from cryonics; the fact that we deconstructed the particles and reconstructed them exactly as they were shouldn't make a difference. But we previously admitted that teleportation is a death sentence and, therefore, we would have to accept that after being cryopreserved, we wouldn't be the same person. Following this line of reasoning, even if we were reconstructed after a split second, we wouldn't be the same. That minute interruption would be our death and the appearance of someone just like us—who, due to the continuity of their memory, would feel that everything worked out and there were no problems with teleportation.

So, what is it that really matters? What maintains our identity and avoids our death? Perhaps it is the continuity in our brain's activity, as we have speculated earlier. However, this is a weak argument since, in principle, there would seem to be no big difference between interrupting the brain activity and suddenly altering it. Our brain activity dramatically changes, for example, if we survive a high-voltage electric shock, a brutal knockout punch, or every time we go to sleep. So, the big question is: Are we constantly dying, reviving as another self that has the illusion of being the same as always because it shares the same memory? Accepting identity, even as a construction of the brain, implies conceiving amazing possibilities and grappling with enigmas for which we do not necessarily have answers.

I would like to cover one more big question before ending this book. Until now, we have reflected on what determines the perception of self and the continuity of our existence, but why does the brain generate this construction? Do we have an evolutionary advantage by being conscious of ourselves? To my understanding, the key is that our sense of identity leads us to ask ourselves about our future and to face death. The fear of death, in turn, leads us to cling to life, even trying to find ways to defy, or at least delay, that which we know is unavoidable.

Toward the end of *Terminator 2*, the T-800 model (played by Arnold Schwarzenegger) doesn't hesitate to destroy his arm to free himself from the machine that has him trapped, and he also doesn't hesitate to plunge into a tank of molten metal to destroy the computer chip in his head.* The supercomputer HAL 9000's reaction in *2001* or the one of android Roy Batty (the replicants' leader) in *Blade Runner* is very different. In these cases, the plot tries to humanize them, so that we don't see them just as machines and instead view them as beings with a sense of identity who worry about their extinction.

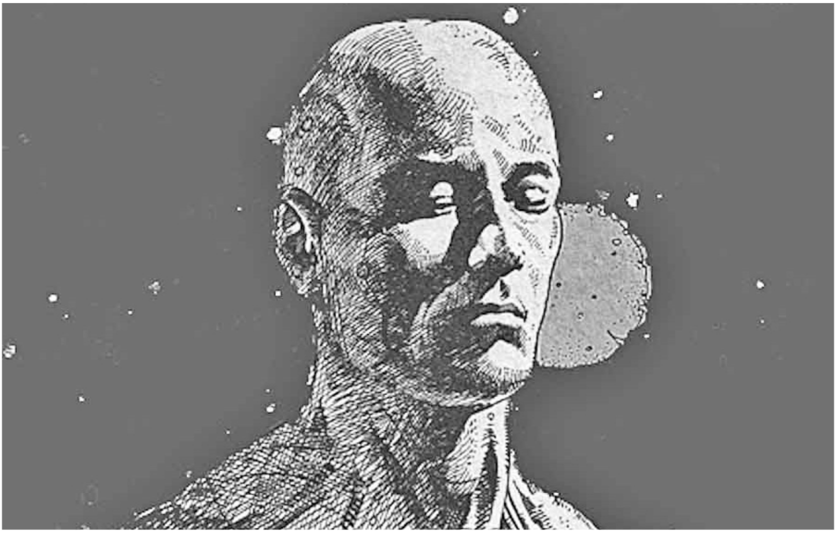
In the first few chapters, we wondered about what differentiates a machine from a human. Besides our unique ability to think, based on the ability to abstract and understand the meaning of things, the feeling of being a person, of having a survival instinct—a conatus—makes us different from machines. In other words, the fact that someone or something fears death shows that there's a being that exists and is conscious of itself.

Picking up the discussion about how to know if a machine is conscious (based on the famous Turing test mentioned in chapter two), we can posit that, like HAL 9000, if one day a computer showed fear of being permanently disconnected (and without this having been specifically programmed, meaning that the behavior is not the consequence of executing an action established in its code), we would probably admit that it is conscious, that it can process and understand its context and have qualia. For example, a computer programmed to recognize faces or visual patterns that is conscious of itself could fear its own end, not wanting to stop feeling or appreciating what it sees.

This also differentiates us from other systems and organisms. With the interaction of billions of users, the internet is extraordinarily complex but it lacks a purpose by itself. The only reason for its existence is a function of the need of its users, just as if the different organs and cells in my body acted on their own accord, with nothing giving them a greater purpose: to preserve my life. If in the future we find a means of interaction that renders the internet obsolete, it would simply disappear without kicking or screaming, because there was never a higher purpose—a consciousness of identity and existence—that led to the preservation of the network itself.

Likewise, consciousness and the fear of death are very limited in inferior beings like insects. A wasp doesn't hesitate for an instant to sacrifice its life when stinging someone, nor does it wonder about what will happen by doing so. In our case, that's not natural; it only happens in extreme circumstances, when the reason that leads us to risk or sacrifice our life is so great that it surpasses our fear of death. It is the consciousness of our own existence that leads us to try to preserve ourselves.

Let's then agree that the self is not a magical or ethereal thing that makes us a person, but rather a construction of the brain. We can ponder whether this existence is real or nominal; however, illusory or not, the feeling of our existence is very strong and it leads us to fear death, to want to prolong our existence, and to even try to become immortal. But beyond our innate reaction, is immortality what we really want?



Gilgamesh, the Immortal. (Image adapted by Candela Insua.)

In Lucho Olivera's comic, mentioned at the start of this chapter, the Martian Utnapishtim keeps his promise to make Gilgamesh immortal and, after finishing the process, he tells the king that when he tires of eternal life he should look for him on the Red Planet. In this ancient era, Gilgamesh asks Utnapishtim how he might reach Mars. The Martian replies that Gilgamesh will have to wait thousands of years, which is insignificant to the now immortal king, until the technological developments of his species allow it. Although Utnapishtim constantly traverses the confines of the universe, Gilgamesh will have all of eternity to find him.

Gilgamesh witnesses the evolution of human civilization and must later spend centuries completely alone after its extinction in a nuclear war. On the one hand, hundreds of years are nothing to him, but he must nonetheless wait each year, each day, each hour until reuniting with Utnapishtim to free himself from the torment of eternal life. Clearly, this is fiction, but it helps us gather strength to face the great taboo and reconsider our attitude toward death. The problem with immortality is that nothing would make sense anymore: Why do something today if we can do it tomorrow, or in a thousand years?

We know we aren't immortal, and yet we try to avoid facing the truth; we fool ourselves so that we can ignore our fate. It is no coincidence that people who have had near-death experiences, are recovering from a serious accident, or are suffering an illness with meager expectations of survival, tend to radically change their lives after asking themselves the questions we mainly try to avoid.

Death gives us an urgency to live. It gives our days meaning. Regardless of how distressing it is to be conscious of our fleeting time, death is life, and instead of trying to defy the inevitable, perhaps it is best to get used to the idea that the present flies by, that the years pass and don't return, and that we have to enjoy the time we have to live because sooner or later everything ends in the sweet sleep.

---

* Unfortunately, Editorial Columba, founded in 1928, closed with the 2001 crisis in Argentina. Our childhood heroes barely survive as tattered issues in dark corners of a few secondhand bookstores and in the memories of their sagas, brightening endless summer afternoons.

* Beyond the possibility of transcending death, science fiction also uses this idea for crew members to go on long interstellar journeys: If we would like to send humans to distant planets, we could keep them in a state of hibernation, perhaps for decades or centuries, and wake them up shortly before they arrive at their destination. This idea is exploited in *2001: A Space Odyssey* and in *Planet of the Apes*, which we have talked about in previous chapters.

* *Open Your Eyes* (*Abre los ojos*) premiered in Madrid in 1997. The movie's success led to a 2001 Hollywood remake titled *Vanilla Sky*, starring Tom Cruise and Penélope Cruz (who also acted in the original film).

* This idea comes from the book *The Prospect of Immortality* (1962) by Robert Ettinger, a physics and math professor in Michigan.

* For cryonics to work (if it will ever work), it is essential that the process begins as soon as possible after the person's death. The longer it takes, the greater the decay, particularly of the brain, given the lack of blood flow, such as in the case of strokes.

** Photo courtesy of Alcor Life Extension Foundation.

* This is a widely used process in neuroscience for in vitro research: slices cut from the brain are preserved in a solution that keeps the neurons alive while their behavior is studied for no more than a few hours through electrophysiological recordings, or recordings of images obtained through specialized microscopes.

* People who are in the cryonics banks used to be millionaires given the procedure's high cost, currently no less than $100,000 to preserve the head, and more than double that amount to preserve the entire body. Oddly enough, it is not clear if these people will retain their fortunes in case they are potentially resuscitated, since if they're legally dead, they have no rights over their belongings.

* The case of the Swedish girl is described in Birgitta S. Romlin et al., "Excellent Outcome with Extracorporeal Membrane Oxygenation After Accidental Profound Hypothermia (13.8°C) and Drowning," *Clinical Care Medicine* 43, no. 11 (2015): e521–e525.

To resuscitate people in these cases, it is essential that hypothermia takes place before hypoxia, meaning that the person's organs freeze and enter a state of hibernation before the lack of oxygen begins to cause cellular death. This is, for example, the case of children who remained submerged in freezing water and drowned after their body temperature dropped dramatically. Paradoxically, the chances of resuscitating a person shortly after they drown are greater if the drowning happens in freezing waters versus warmer waters. To review these cases, see: Peter Paal et al., "Accidental Hypothermia–An Update," *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 24 (2016): 111.

* It is still up for debate whether George Mallory and Andrew Irvine, his climbing partner, were the first to summit Everest before dying.

** There are no reports on this because in such a critical situation the last thing on the doctors' minds is to do an electroencephalogram.

* To simplify, we'll limit ourselves to copying the brain without trying to also reproduce the body. The brain would be isolated from the outside, but in principle it would maintain our memories—our self—and it would be a way of preserving ourselves.

** Considering only neurons that connect among each other is also a huge simplification. First, we have to distinguish between anatomic connectivity (the one we see in a microscope) and functional connectivity (the connections that are active). Second, neurons receive connections from other neurons in their dendrites, and the exact location of these connections within the dendrite tree influences their responses. For more information see, for example, Tiago Branco, Beverley A. Clark, and Michael Häusser, "Dendritic Discrimination of Temporal Input Sequences in Cortical Neurons," *Science* 329, no. 5999 (2010): 1671–1675.

* It already shows up in Fritz Lang's *Metropolis*, the first movie in this genre (released in 1927), where corrupt city rulers replace an influential woman, Maria, with a robot in an attempt to co-opt unwitting factory workers.

** Curiously enough, *Star Trek's* teleportation idea—one of sci-fi's most famous resources—comes from a ridiculous practical limitation. The *Enterprise*'s design made it difficult and expensive to film scenes of the craft landing on different planets. The solution was to teleport the crew so that the ship would always remain in space.

For an interesting discussion on the physics principles involved in teleportation, see chapter five in *The Physics of Star Trek*, by American physicist Lawrence Krauss (New York: Basic Books, 1995).

* It is worth noting that in order to avoid the problems that come with transporting particles at great speeds, the only thing that is transmitted is the information on how the particles are assembled, replicating our body with available particles at the final destination. Pure functionalism.

* John Perry, *A Dialogue on Personal Identity and Immortality* (Indianapolis: Hackett Publishing, 1978). Gretchen Weirob, a philosopher who lies on her deathbed, discusses with a minister and a former student the concrete possibility of life after death. Through this dialogue, Perry critically reviews different philosophical and religious positions on personal identity and its continuity.

* Curiously, in the movie's original script, Sarah Connor asks the Terminator if he is afraid, right before dying, to which he answers yes. However, this dialogue does not appear in the movie, fixing in the last minute what would've been quite incoherent, since by presenting the Terminator as an emotionless machine, it would not make sense for him to fear death.

EPILOGUE

The Dawn of a New Philosophy

René Descartes revolutionized philosophy by systematically doubting all preestablished knowledge, and he was even able to challenge something as immediate, although not at all obvious, as the existence of reality. To revisit his age-old query: How do we know that everything we see around us isn't just a colossal deception whipped up by an evil genius? The impossibility of simulating the extremely complex interactions we have with the external world and, as Wittgenstein argued, the impossibility of even acquiring the tiniest bit of knowledge without having someone to interact with, leads us to conclude that reality exists. Although reality isn't necessarily what it appears to be, since we don't have access to reality itself—quoting Kant, *die Dinge an sich*—but to our subjective interpretation of it. In other words, the perception we have of reality is a construction of the brain.

We are currently experiencing a similar revolution to the one Descartes unleashed, which, like before, is once again shaking the foundations of philosophy. Descartes posited a division between mind and matter, but scientists have set aside Cartesian dualism, and we now mainly support materialism. We no longer see the mind as a separate entity from the brain, but as the latter's activity. Mind and brain are one and the same, just like the temperature and the kinetic energy of the microscopic movement of molecules—two different ways of describing the same phenomenon. However, adopting materialism gives rise to other apparently unsolvable issues that have been keeping both philosophers and neuroscientists up at night for a while. Perhaps the biggest one of all has been the "hard problem," which is related to consciousness. If the mind is nothing more than the firing of neurons, how can electric impulses in my brain make me feel and be aware of, for example, the smell of jasmine?

The sci-fi movies we have discussed were, to a degree, a springboard that led us to raise these questions—questions that great thinkers such as Aristotle, Descartes, Locke, and Hume, among others, have been asking for centuries; questions whose answers we are now beginning to unlock. The key is always the same: understanding that mental processes are a construction of the brain.

Does consciousness exist? Of course it does, though not as an ethereal and mysterious entity, but as the coordinated activity of groups of neurons. How can we then explain the feeling of smelling jasmine? Easy. When we smell jasmine, we activate groups of neurons that identify this particular smell, which in turn activate other groups of neurons coding related feelings, experiences, and emotions: the color and texture of its petals, what my grandmother's jasmine plant looked like, the delight she felt when smelling its fragrance, and many other things that when evoked, consciously or unconsciously, generate the subjective experience that I have when I smell jasmine—in other words, qualia.

But, how do we explain the feeling of pain? It is clear that pain is a defense mechanism, for example, to prompt us to remove our hand from a hot fire. Still, how is the feeling that something hurts generated from the firing of certain groups of neurons? Why does the activation of those neurons make us tremble with pain, while others give us pleasure? The answer likely comes from the connectivity of those neurons and the group of activations they trigger, but we are still far from understanding how these processes actually work. However, departing from Leibniz's view—who famously argued that if cognitive processes were like a large windmill, when entering it, we would only see pieces interacting with each other but nothing that could explain these processes—we currently believe that by knowing how different mechanisms work in the brain, we can, in fact, begin to explain cognitive processes and, particularly, something as elusive as consciousness.

Another big issue that keeps philosophers (and Western religious scholars) up at night is free will, especially after the scientific revolution of the seventeenth century and the rise of determinism. How can we be free if everything we are about to do is already predetermined? Scholastic philosophers, notably Thomas Aquinas, asked themselves how we could be free if God knows beforehand every one of our actions (and, on the other hand, what's left of our freedom if we can't choose to sin).

The idea of an autonomous mind that makes decisions using its own will becomes unsustainable from a scientific point of view. Free will is another construction of the brain, generated by consciously anticipating actions that are about to be executed. But these actions are generated through the firing of neurons and are, therefore, predetermined. This gives rise to devastating consequences that prompt us to reconsider the responsibility we have for our actions. Furthermore, a lack of free will seems to make our lives meaningless, since everything we might decide in the future, every single eventual success and failure, has already been predetermined. However, chaos theory and the resulting unpredictability of the universe and, specifically, of the activity of our brain (and therefore of our behavior) saves us from this great philosophical dilemma. Our actions are unpredictable and, in practice, it doesn't really matter that the future is predetermined if no one can see it.

We are currently experiencing a revolution in artificial intelligence, and sophisticated algorithms are already surpassing humans at endless tasks. How then do we know that a machine isn't intelligent or capable of thinking? The question is far from trivial. We assume that other people think like us because they show similar behaviors (which is the basis of the theory of mind; of "putting ourselves in someone else's shoes" to interact socially). But could we say the same of a computer? What would a computer or an android have to do to show us that it is intelligent? These are the questions that Turing asked himself, giving way to his famous test, which, however, is quite controversial when trying to put it into practice.

Nevertheless, we are still far from being surpassed or even paralleled by machines; they can excel in specific tasks, but they lack general intelligence. And I think this is key to determining if a machine can think: What matters is not that it can solve a certain problem, but whether it can solve problems that it hasn't been trained for, showing understanding and the power to transfer knowledge based on analogies and inferences. This is something we do constantly, when we face situations we have never experienced before, scenarios in which we know how to react using common sense, making inferences through past experiences. We don't need, as does a machine, a zillion examples and specific practice to successfully deal with the new problems we face so often in our lives. And on many occasions, we can't really judge if something is right or wrong; there's no rule for what to do, and not even a clear outcome—like winning or losing a chess game—that we can use to tune our behavior. Rather, we have a sense of a given situation, we understand what's going on, and somehow decide what's best. But our ability to understand is based on extracting meaning from things, processing that which is essential, discarding infinite irrelevant information. This is precisely the basis of brain function, the result of millions of years of evolution. And it wouldn't surprise me if the two outmost challenges of AI—general intelligence and consciousness—are somehow linked; if the ability to develop general intelligence, a capacity for making sense of things, would lead to the development of a sense of self, the so-far elusive problem of how to make a machine aware of its own existence.

The key to brain function is that it stores very little information because it mainly uses its resources to understand. This is the big difference with how a computer works. A computer can reliably store an enormous amount of information, but it doesn't understand its content. The brain, in contrast, deals with little information because it focuses on the essential and infers the rest to attribute meaning to things. Information is not simply stored in the brain—it is constructed. And it is tempting to argue that this active construction process, so far lacking in machines, is key for general intelligence, and may also be the crucial feature that gives rise to consciousness. However, we do want machines to accurately store and process information; we don't want them to construct this information and potentially fail when making incorrect assumptions, as we, for example, do when we experience visual illusions or false memories. This is not the way we build our computers, and who knows what will happen if one day we do so.

Delving further into this discussion, we could also ask ourselves: Why don't other animals have our intelligence? A chimpanzee's brain isn't that different from that of a human. However, there's a noticeable difference that stems from the use of language and everything it facilitates. Language allows us to communicate, transmit knowledge, discuss past experiences, and plan the future. But beyond these unique possibilities, I think the greatest benefit of language is the development of a more conceptual and abstract thought, setting aside the immediate details and peculiarities of things. As Borges said: "Every noun is an abbreviation," and "to think is to forget differences, to generalize, to abstract." Furthermore, Borges argued (in "Tlön, Uqbar, Orbis Tertius") that in a world without nouns, science and even knowledge would be impossible. So, I dare speculate that, after tens of thousands of years of evolution, one of the consequences of language is the development of concept cells, like those I discovered a few years ago doing experiments with humans—neurons that, incidentally, have yet to be found in animals and may very well be the cornerstone of our capacity for conceptual and abstract thought.

The representations given by concept cells are the basis of our memory, since the memory of our experiences derives from storing relatively few associated concepts (for example, remembering that I was at a café with a friend talking about a particular subject). The rest is mainly lost in oblivion, allowing us to focus on the essential to generate associations and contextualize each new piece of knowledge—something we usually do in our sleep, where we probe even the most bizarre associations. This ability to forget details and focus on the meaning of things is precisely the basis of our intelligence and creativity, our capacity to make associations between initially disparate concepts to understand, for example, that an apple that falls to the ground and the moon that orbits in the firmament respond to the same phenomenon: the law of gravity.

Helmholtz maintained that perception is based on processing very little information and making unconscious inferences. We see very little and assume the rest. And like perception, memory is also a construction of the brain. Our feeling of seeing an object is based on identifying it through just a few features. Likewise, the feeling of remembering our past as if it were a movie is based on creating a coherent story by recalling very few events, filling in the blanks using inferences based on common sense. Such constructions of the brain underlie our sense of identity and the feeling that our bodies belong to us. The hand I see typing on the keyboard is my hand because it responds to commands dictated by my brain. If it didn't, I would feel that it was someone else's hand, as with the alien hand syndrome. Conversely, I would consider as my own a synthetic hand which, using the latest achievements in neuroprosthetics, responds reliably to commands from my brain.

The sense of identity, of continuing to be the same person, is a major philosophical dilemma and, as Locke argued, does not come from our body but from memory continuity. I am the person who not so long ago decided to write this book, the one who studied physics at the University of Buenos Aires, and was blown away by movies such as *2001, Blade Runner, The Matrix*, and *Until the End of the World*. But, surprisingly, this story we create—our story—is a dynamic process that is constantly evolving, because each time we consciously access a memory we are changing it.

So, what's left of the self? How can it be that what I am most certain about, my own existence, is generated by something as unstable and malleable as my memory? The notion of self is perhaps one of the most elusive and controversial concepts in philosophy. To Hume it is nothing more than a bunch of sensations, but to others, the self, the feeling of being a person, is undeniable. In my view, the self exists and it is the most elaborate construction of the brain, giving entity and identity to that bunch of sensations. (Moreover, we saw that there is a representation of the self in the brain, when we found neurons in patients that responded to pictures of themselves.)

By concluding that the construction of the self is based on our memory, and knowing that memory is, in turn, triggered by the activity of neurons and their connections, we then wonder if we could preserve our identity—in a clone or supercomputer—and somehow become immortal. That's twenty-first-century philosophy. Therein lies the revolution where the most challenging topics in philosophy—identity, consciousness, free will, and animal and machine intelligence among them—now seem approachable, and we can even face something as elusive as immortality as a biological and not just a metaphysical problem.

To defeat death is perhaps our most compelling challenge; however, we saw that even if we could make an exact copy of our brain, we wouldn't be what was generated. We would create a person with our same memories, our same feeling of identity—of having a self—but that person wouldn't be us. For a third person, the difference between us and our clone would be unrecognizable, but we wouldn't see through the clone's eyes or feel what he or she feels. When conceiving the possibility of cloning or teleportation, like in science fiction, we then realize that the self doesn't necessarily have to be unique.

Fear of death is the most natural instinct and comes from the construction of the self, because if there is no self that is going to die, the fear is lost. That is, in my opinion, the fundamental consequence of our brain's creation of a self: Spinoza's conatus, the need to try to preserve ourselves. Computers and other complex systems like the internet are not aware of themselves (at least to our knowledge). The internet lacks a purpose, a sense of self-preservation; its raison d'être comes from the desires of billions of users, just as if each one of the billions of neurons in our brain advocated for their own well-being and existence, but without fostering the self, the person's well-being. The internet could disappear tomorrow without a problem, if its users found another more efficient way to share and exchange information, just like the cells in my liver wouldn't mind (or be conscious of) their extinction if I have a transplant. But it would be a very different situation if the neurons in my brain could somehow manage to reconnect themselves and create another person, because that would determine my death, the disappearance of the greater purpose, and I would certainly do everything in my power to prevent this from happening.

Therefore, consciousness provides us with a preservation instinct. Returning to the discussion about Turing's test, a machine could prove that it is intelligent if, like HAL 9000 in *2001* or Skynet in *Terminator*, it tried at all costs to avoid being disconnected—of course, assuming this is not something that has been preprogrammed and instead results as a consequence of its self-awareness. But how could we induce this higher purpose in a machine? How could we create the awakening of its consciousness and its fear of death? Would the induction of general intelligence do that? This is without a doubt one of the most fascinating and recurring questions in science fiction and twenty-first-century philosophy.

We have begun to see how the brain generates a construction of the self, but what determines its continuity? What makes me the same when I wake up each morning but not so if I'm revived after years of being frozen? Are we perhaps constantly dying and resuscitating as another self that believes it is the same as always? Is our dreaded death nothing more than a disruption in this chain of rebirths?

In recent years, science has made enormous advances, but questions like these are still standing, unanswered. Despite how frustrating this may seem, it is the Holy Grail for scientists, that which leads us to tirelessly work to satisfy our curiosity, enjoying each moment in the search for answers that we may never find. However, what is undeniable is that, like never before in the history of humanity, we now have methods and tools that we did not even imagine a few years ago, placing us in a privileged position to discuss these questions one-on-one with the best thinkers of all times and scientifically tackle the greatest philosophical challenges.

ACKNOWLEDGMENTS

As a scientist, I stick to facts, and my typical writings, in the form of research articles, describe the results of experiments. That’s the norm, but sometimes I feel it’s not enough and would like to go beyond and ask “what if . . .”, even if I can’t conceive of how to tackle the problem scientifically. In this book I tried to explore that and let my imagination go. An unexpected result of this is that thinking about these issues has triggered a radical change in my research. Before, I focused on studying the neurons I discovered a few years ago, and on showing how these neurons form and encode memories. Now, I’m more interested in trying to understand what makes us human and the experiments we could potentially do to figure out what it is in our brain that differentiates us from machines and other animals—the topic of so many sci-fi stories. So, oddly enough, I first have to thank many people I’ve never met: science fiction writers and directors who have, since I was a kid, carried me away and let my mind travel to imaginary worlds and situations that we don’t encounter in real life.

What was science fiction just a few years ago is perhaps that to which I’m going to dedicate the rest of my scientific career. There’s nothing that interests me more these days—the ever more blurred distinction between science and science fiction. It was a lot of fun to write about this, and was even more fun to discuss which of these extraordinary feats are now becoming reality. I felt this was a fascinating story to tell, and if I succeeded in doing it well, it is to a large extent thanks to all the careful editorial work of Cecilia, Laurel, Alexa, and Scott. Each chapter illustration also prompts us to jump straight to its content, and I’m thankful to Candela Insua for that.

Writing this book was a joy but at times also hard. It’s not that easy to write about the lack of free will, the fear of death, or dreams of immortality. And this is not just because these are difficult topics to cover, it is because they often left me with a feeling of emptiness—the bitter sensation that everything seems meaningless and ephemeral. Why should we care about anything if death is unavoidable and everything we manage to do and achieve will be wiped out in time? Should we end up like Diogenes, in ancient Greece, living in a barrel and not caring about anything?

Believe me, I struggle as much as you do. But if there’s something I have learned, it is that the worst possible attitude is to be in denial—to behave as if these issues don’t exist. I’m afraid I can’t offer a general answer to these questions; I think it’s something very personal. In my particular case, I’m lucky enough to have a wonderful wife and three adorable children. So, what’s the purpose? Why did I even bother to write this book? Well

. . . in this particular matter I have a very good answer: I wrote it for them.

NOTES

Chapter 1

1 Marvin Minsky, *The Society of Mind* (New York: Touchstone/Simon & Schuster, Inc., 1988).

2 Minsky, *The Society of Mind*, 20.

3 Minsky, *The Society of Mind*, 326.

4 Translated from *Obras completas de Platón*, vol. 5 (Madrid: Edición Patricio de Azcárate, 1871).

5 Translated from *Meditaciones metafísicas con objeciones y respuestas*, sexta Meditación (Madrid: Edición de Vidal Peña, Alfaguara, 1977).

6 For more information about Phineas Gage, see, for example, Malcolm Macmillan, "Phineas Gage—Unraveling the Myth," *The Psychologist* 21 (2008): 828–831.

7 H.M. was studied for thirty years by the Canadian psychologist Brenda Milner. For more information, see the first scientific report Milner wrote together with Scoville about H.M.: William Scoville and Brenda Milner, "Loss of Recent Memory After Bilateral Hippocampal Lesion," *Journal of Neurology, Neurosurgery, and Psychiatry* 20 (1957): 11–21.

8 Gilbert Ryle, *The Concept of Mind* (London, New York: Hutchinson, 1949).

9 Francis Crick, *The Astonishing Hypothesis: The Scientific Search for Soul* (New York: Scribner, 1994).

10 Rodrigo Quian Quiroga, *Borges and Memory: Encounters with the Human Brain* (Cambridge, MA: MIT Press, 2012).

11 For more information about the algorithm, see Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* 25, no. 2 (2012): 1097–1105.

12 For a compilation of the performances of different algorithms in the last few years and how they compare with human performance, see Olga Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115, no. 3 (2015): 211–252.

13 For more information, see Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521, no. 7553 (2015): 436–444.

14 LeCun et al., "Deep Learning."

15 Christof Koch, "How the Computer Beat the Go Master," *Scientific American*, March 19, 2016.

16 David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529, no. 7587 (2016): 484–489.

17 Asimov's Rules of Robotics appeared in his 1942 short story "Runaround," which was included in his *I, Robot* collection published in 1950.

Chapter 2

1 David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996).

2 Semir Zeki, "Parallel Processing, Asynchronous Perception, and a Distributed System of Consciousness in Vision," *The Neuroscientist* 4, no. 5 (1998): 365–372.

3 Daniel C. Dennett, *Consciousness Explained* (London: Penguin, 1993), 37.

4 Francis Crick and Christof Koch, "The Problem of Consciousness," *Scientific American* 267, no. 3 (1992): 152–159. Crick and Koch's article was republished in the same journal in 2002.

5 For more information, see Randolph Blake and Nikos K. Logothetis, "Visual Competition," *Nature Reviews Neuroscience* 3, no. 1 (2002): 13–21.

6For more information see, for example, Nikos K. Logothetis, “Single Units and Conscious Vision,” *Philosophical Transactions of the Royal Society B-Biological Sciences* (1998): 1801–1818.

7For more information see, for example, Larry R. Squire and Stuart Zola-Morgan, “The Medial Temporal Lobe Memory System,” *Science* 253, no. 5026 (1991): 1380–1386.

8For more information about the electrodes, the process involved in recording neurons, and further technical details see, for example, Hernan G. Rey, Matias J. Ison, Carlos Pedreira, Antonio Valentin, Gonzalo Alarcon, Richard Selway, Mark P. Richardson, and Rodrigo Quian Quiroga, “Single-Cell Recordings in the Human Medial Temporal Lobe,” *Journal of Anatomy* 227, no. 4 (2015): 394–408.

9Others who participated in these experiments include Gabriel Kreiman, Leila Reddy, Alexander Kraskov, Florian Mormann, Christof Koch, and Itzhak Fried (the neurosurgeon who established these types of recordings at UCLA).

10Rodrigo Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried, “Invariant Visual Representation by Single Neurons in the Human Brain,” *Nature* 435, no. 7045 (2005): 1102–1107.

11For more examples, see Rodrigo Quian Quiroga, Alexander Kraskov, Christof Koch, and Itzhak Fried, “Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain,” *Current Biology* 19, no. 15 (2009): 1308–1313.

12For a summary of the results we obtained with these neurons, see Rodrigo Quian Quiroga, “Concept Cells: The Building Blocks of Declarative Memory Functions,” *Nature Reviews Neuroscience* 13, no. 8 (2012): 587–597.

13The results of this experiment are described in Rodrigo Quian Quiroga, Roy Mukamel, Eve A. Isham, Raphael Malach, and Itzhak Fried, “Human Single-Neuron Responses at the Threshold of Conscious Recognition,” *Proceedings of the National Academy of Sciences* 105, no. 9 (2008): 3599–3604.

14The results of this experiment are described in: Rodrigo Quian Quiroga, Alexander Kraskov, Florian Mormann, Itzhak Fried, and Christof Koch, “Single-Cell Responses to Face Adaptation in the Human Medial Temporal Lobe,” *Neuron* 84, no. 2 (2014): 363–369.

15Chapter four in *Borges and Memory* (cited in chapter one) is devoted to a description of H.M. The first report about the H.M. case is by William Scoville and Brenda Milner, “Loss of Recent Memory After Bilateral Hippocampal Lesion,” *Journal of Neurology, Neurosurgery and Psychiatry* 20, no. 1 (1957): 11–21.

A concise description of his contribution to science is given in Larry R. Squire, “The Legacy of Patient H. M. for Neuroscience,” *Neuron* 61, no. 1 (2009): 6–9.

16This is what we could call *perceptual* and *contextual awareness*. For more information, see:

- Joaquin Navajas, Hernan G. Rey, and Rodrigo Quian Quiroga, “Perceptual and Contextual Awareness: Methodological Considerations in the Search for the Neural Correlates of Consciousness,” *Frontiers in Psycho*logy 5 (2014): 959.
- Rodrigo Quian Quiroga, “Neuronal Codes for Visual Perception and Memory,” *Neuropsychologia* 83 (2016): 227–241.

17Turing proposed his famous test in Alan M. Turing, “Computing Machinery and Intelligence,” *Mind* 59, no. 236 (1950): 433–460.

18For more information, see www.square-bear.co.uk/mitsuku/home.htm

19Rashid’s demonstration can be seen on YouTube, where it has more than a million views: “Speech Recognition Breakthrough for the Spoken, Translated Word.” Posted by Microsoft Research, November 8, 2012. www.youtube.com/watch?v=Nu-nlQqFCKg

20For a critical discussion, see Searle’s original article and the following commentary by different authors in John R. Searle, “Minds, Brains, and Programs,” *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–457.

Chapter 3

1Michel de Montaigne, *An Apology for Raymond Sebond* (Translated from *Apología de Raimundo Sabunde*), 386, 388.

2René Descartes, *A Discourse on the Method*, V (1637).

3Montaigne, *An Apology for Raymond Sebond.*

4Montaigne, *An Apology for Raymond Sebond.*

5Plutarch, *Moralia* $973_{\text{c–e}}$; Montaigne, *An Apology for Raymond Sebond.*

6For an informational discussion on the experiments done on jays, see V. Morell, “Nicky and the Jays,” *Science* 315, no. 5815 (2007): 1074–1075. And for a more detailed and technical discussion, see Uri

Grodzinski and Nicola S. Clayton, “Problems Faced by Food-Caching Corvids and the Evolution of Cognitive Solutions,” *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365 (2010): 977–987.

7This work’s results were published in Christopher Krupenye et al., “Great Apes Anticipate That Other Individuals Will Act According to False Beliefs,” *Science* 354, no. 6308 (2016): 110–114.

8For a detailed analysis of various deceitful behaviors in higher primates, see Richard W. Byrne and Andrew Whiten, “Cognitive Evolution in Primates: Evidence from Tactical Deception,” *Man* 27, no. 3 (1992): 609–627.

9For more details see, for example, Gordon Gallup, Jr., “Chimpanzees: Self-Recognition,” *Science* 167, no. 3912 (1970): 86–87.

10For more details on these and other experiments with monkeys see, for example, chapter four of José Luis Díaz, *La conciencia viviente* (Fondo de Culture Económica, 2007), or chapter five of Carl Sagan, *The Dragons of Eden* (New York: Ballantine Books, 1986).

11John Locke, *An Essay Concerning Human Understanding* (Book II, Ch. 11, Sec. 10), 1689.

12See, for example, Pavel M. Itskov, Ekaterina Vinnik, and Mathew E. Diamond, “Hippocampal Representation of Touch-Guided Behavior in Rats: Persistent and Independent Traces of Stimulus and Reward Location,” *PLOS One* 6, no. 1 (2011): e16462.

Chapter 4

1René Descartes, *Meditations* (First Meditation) (cited in chapter one).

2For a detailed description on the subject see, for example, Anthony Kenny, *A Brief History of Western Philosophy* (New Jersey: Wiley, 1998); Bertrand Russell, *History of Western Philosophy* (Abingdon-on-Thames: Routledge Classics, [1946] 2004); John Hospers, *An Introduction to Philosophical Analysis* (Abingdon-on-Thames: Routledge, 1956). The online version of the *Stanford Encyclopedia of Philosophy* also offers a detailed and clear description of idealism.

3C. Daniel Salzman, Kenneth H. Britten, and William T. Newsome, “Cortical Microstimulation Influences Perceptual Judgement of Motion Direction,” *Nature* 346, no. 6280 (1990): 174–177.

4Seyed-Reza Afraz, Roozbeh Kiani, and Hossein Esteky, “Microstimulation of Inferotemporal Cortex Influences Face Categorization,” *Nature* 442, no. 7103 (2006): 692–695.

5Wilder Penfield, *The Mystery of the Mind* (Princeton: Princeton University Press, 1975).

6For more details, see: Itzhak Fried et al., “Electric Current Stimulates Laughter,” *Nature* 391, no. 6668 (1998): 650.

7Daniel C. Dennett, *Consciousness Explained* (London: Penguin, 1993), 4.

8For a more detailed and quantitative description on how much we see, see chapter two of *The Forgetting Machine* (cited in chapter two). Chapter three in that book also describes the processes implemented by the brain to process visual information based on unconscious inferences.

9For more information, see my previously cited essay on “Magic and Cognitive Neuroscience,” *Current Biology* 26, no. 10 (2016): R387–R407.

10For a more detailed description of these results, see chapter five of Frederic Bartlett’s *Remembering* (Cambridge: Cambridge University Press, 1932). Chapter seven also describes similar results with other stories.

11Jorge Luis Borges, *Fictions* (cited in chapter three).

Chapter 5

1Charlie Dunbar Broad, *The Mind and Its Place in Nature* (London: Kegan Paul, Trench, Trubner & Co, 1925), 71.

2For more information, see Frank Jackson, “Epiphenomenal Qualia,” *Philosophical Quarterly* 32, no. 127 (1982): 127–136.

3Nagel describes this idea in Thomas Nagel, “What Is It Like to Be a Bat?,” *The Philosophical Review* 83, no. 4 (1974): 435–450.

4Daniel C. Dennett, *Consciousness Explained* (London: Penguin, 1993), 398–401.

5Dennett, *Consciousness Explained*, 441–448.

6For a more detailed description of these algorithms, see Rodrigo Quian Quiroga and Stefano Panzeri, “Extracting Information from Neural Populations: Information Theory and Decoding Approaches,” *Nature Reviews Neuroscience* 10, no. 3 (2009): 173–185.

7Rodrigo Quian Quiroga, Leila Reddy, Christof Koch, and Itzhak Fried, “Decoding Visual Inputs from

Multiple Neurons in the Human Temporal Lobe," *Journal of Neurophysiology* 98, no. 4 (2007): 1997–2007.

8For more information, see Moran Cerf et al., "On-line, Voluntary Control of Human Temporal Lobe Neurons," *Nature* 467, no. 7319 (2010): 1104–1108.

9Adrian M. Owen et al., "Detecting Awareness in the Vegetative State," *Science* 313, no. 5792 (2006): 1402.

10Martin M. Monti et al., "Willful Modulation of Brain Activity in Disorders of Consciousness," *The New England Journal of Medicine* 362, no. 7 (2010): 579–589.

11These results are described in Uri Hasson et al., "Intersubject Synchronization of Cortical Activity During Natural Vision," *Science* 303, no. 5664 (2004): 1634–1640.

12James V. Haxby et al., "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex," *Science* 293, no. 5539 (2001): 2425–2430.

13The analysis method developed by Haxby and his colleagues has given way to endless scientific works in the last few years. For a review of these works see, for example, John-Dylan Haynes and Geraint Rees, "Decoding Mental States from Brain Activity in Humans," *Nature Reviews Neuroscience*, no. 7 (2006): 523–534. For a more detailed description of the methods, see John-Dylan Haynes, "A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives," *Neuron* 87, no. 2 (2015): 257–270.

14See, for example, Kendrick N. Kay et al., "Identifying Natural Images from Human Brain Activity," *Nature* 452, no. 7185 (2008): 352–355.

15For more details, see Shinji Nishimoto et al., "Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies," *Current Biology* 21, no. 19 (2011): 1641–1646.

16Kathleen M. O'Craven and Nancy Kanwisher, "Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions," *Journal of Cognitive Neuroscience* 12, no. 6 (2000): 1013–1023.

17T. Horikawa et al., "Neural Decoding of Visual Imagery During Sleep," *Science* 340, no. 6132 (2013): 639–642.

18Doris Y. Tsao et al., "A Cortical Region Consisting Entirely of Face-Selective Cells," *Science* 311, no. 5761 (2006): 670–674.

19Winrich A. Freiwald and Doris Y. Tsao, "Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System," *Science* 330, no. 6005 (2010): 845–851.

20Le Chang and Doris Y. Tsao, "The Code for Facial Identity in the Primate Brain," *Cell* 169, no. 6 (2017): 1013–1028. For a short description of these results and their implications, see Rodrigo Quian Quiroga, "How Do We Recognize a Face?" *Cell* 169, no. 6 (2017): 975–977.

Chapter 6

1F. W. Truscott and F. L. Emory, trans. *A Philosophical Essay on Probabilities* (New York: John Wiley & Sons, 1902), 4.

2Benjamin Libet et al., "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act," *Brain* 106, no. 3 (1983): 623–642.

3For a critical discussion of Libet's experiment, see Benjamin Libet, "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action," *Behavioral and Brain Sciences* 8, no. 4 (1985): 528–566.

4Itzhak Fried et al., "Functional Organization of Human Supplementary Motor Cortex Studied by Electrical Stimulation," *The Journal of Neuroscience* 11, no. 11 (1991): 3656–3666.

5Joaquim P. Brasil-Neto et al., "Focal Transcranial Magnetic Stimulation and Response Bias in a Forced-Choice Task," *Journal of Neurology, Neurosurgery, and Psychiatry* 55, no. 10 (1992): 964–966.

6Sergio Della Sala, a neurologist specializing in alien hand syndrome, often refers to Cortázar and Peter Sellers in his publications. See, for example, Sergio Della Sala, "The Anarchic Hand," *The Psychologist* 18, no. 10 (2005): 606–609.

7For a description of these cases see, for example, D. H. Geschwind et al., "Alien Hand Syndrome: Interhemispheric Motor Disconnection Due to a Lesion in the Midbody of the Corpus Callosum," *Neurology* 45, no. 4 (1995): 802–808; and Gary Goldberg, Nathaniel H. Mayer, and Joseph U. Toglia, "Medial Frontal Cortex Infarction and the Alien Hand Sign," *Archives of Neurology* 38, no. 11 (1981): 683–686.

8Sarah-Jayne Blakemore, Daniel M. Wolpert, and Christopher D. Frith, "Abnormalities in the Awareness of Action," *Trends in Cognitive Sciences* 6, no. 6 (2002): 237–242.

9P. K. McGuire, R. M. Murray, G. M. S. Shah,, “Increased Blood Flow in Broca’s Area During Auditory Hallucinations in Schizophrenia,” *Lancet* 342, no. 8873 (1993): 705–706.

10Sean Spence, “Free Will in the Light of Neuropsychiatry,” *Philosophy, Psychiatry, and Psychology* 3, no. 2 (1996): 75–90.

11Sarah-Jayne Blakemore, Christopher D. Frith, and Daniel M. Wolpert, “Spatio-Temporal Prediction Modulates the Perception of Self-Produced Stimuli,” *Journal of Cognitive Neuroscience* 11, no. 5 (1999): 551–559.

12These results are described in Kathleen D. Vohs and Jonathan W. Schooler, “The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating,” *Psychological Science* 19, no. 1 (2008): 49–54.

13Jim Fallon’s search to explain if he was in effect a psychopath is the plot of his book *The Psychopath Inside* (New York: Penguin Group, 2013).

14Michael S. Gazzaniga, *Who’s in Charge?* (London: Robinson, 2012).

15For more information on these cases and statistics, see the May 23, 2016, report “Machine Bias” on the *ProPublica* portal.

16Petter Johansson et al., “Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task,” *Science* 310, no. 5745 (2005): 116–119.

17This study is described in Lars Hall et al., “How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions,” *PLOS* One 8, no. 4 (2013): e60554.

Chapter 7

1For more information, see:

- Miguel A. L. Nicolelis, “Actions from Thoughts,” *Nature* 409, no. 6818 (2001): 403–407.
- John P. Donoghue, “Bridging the Brain to the World: A Perspective on Neural Interface Systems,” *Neuron* 60, no. 3 (2008): 511–521.
- Bijan Pesaran, Sam Musallam, and Richard A. Andersen, “Cognitive Neural Prosthetics,” *Current Biology* 16, no. 3 (2006): R77–R80.

2See, for example, Rodrigo Quian Quiroga, Lawrence H. Snyder, P. Aaron Batista, He Cui, and Richard A. Andersen, “Movement Intention Is Better Predicted Than Attention in the Posterior Parietal Cortex,” *Journal of Neuroscience* 26, no. 13 (2006): 3615–3620.

3In this case, neural recordings were taken from the motor cortex. For more details, see Meel Velliste et al., “Cortical Control of a Prosthetic Arm for Self-Feeding,” *Nature* 453, no. 7198 (2008): 1098–1101.

4Among others, an article in the *New York Times*, “Monkeys Think, Moving Artificial Arm as Own,” published May 29, 2008.

5For more information, see Leigh R. Hochberg et al., “Reach and Grasp by People with Tetraplegia Using a Neurally Controlled Robotic Arm,” *Nature* 485, no. 7398 (2012): 372–375; and Tyson Aflalo et al., “Decoding Motor Imagery from the Posterior Parietal Cortex of a Tetraplegic Human,” *Science* 348, no. 6237 (2015): 906–910.

6For more information, see Hochberg et al., “Reach and Grasp by People with Tetraplegia Using a Neurally Controlled Robotic Arm”; Aflalo et al., “Decoding Motor Imagery from the Posterior Parietal Cortex of a Tetraplegic Human.”

7For more information, see Beata Jarosiewicz et al., “Functional Network Reorganization During Learning in a Brain-Computer Interface Paradigm,” *Proceedings of the National Academy of Science of USA* 105, no. 49 (2008): 19486–19491.

8Plutarch, *Parallel Lives*, “Life of Theseus,” XXIII, 1.

9John Locke, *An Essay Concerning Human Understanding*, Book II, Chapter XXVII, sec. 15 (1689).

10Matthew Botvinick and Jonathan Cohen, “Rubber Hands ‘Feel’ Touch That Eyes See,” *Nature* 391, no. 6669 (1998): 756.

11For more information, see H. Henrik Ehrsson, “The Experimental Induction of Out-of-Body Experiences,” *Science* 317, no. 5841 (2007): 1048; and Ed Yong, “Master of Illusion,” *Nature* 480, no. 7376 (2011): 168–170.

12Marco Capogrosso et al., “A Brain-Spine Interface Alleviating Gait Deficits After Spinal Cord Injury in Primates,” *Nature* 539, no. 7628 (2016): 284–288.

13Andrew Jackson, “Spinal-Cord Injury: Neural Interfaces Take Another Step Forward,” *Nature* 539, no. 7628 (2016): 177–178.

14C. Ethier et al., “Restoration of Grasp Following Paralysis Through Brain-Controlled Stimulation of Muscles,” *Nature* 485, no. 7398 (2012): 368–371.

15Chad E. Bouton et al., “Restoring Cortical Control of Functional Movement in a Human with Quadriplegia,” *Nature* 533, no. 7602 (2016): 247–250.

16Miguel Pais-Vieira et al., “A Brain-to-Brain Interface for Real-time Sharing of Sensorimotor Information,” *Scientific Reports* 3 (2013): 1319.

17For more information, see:

- J. O. Mills, A. Jalil, and P. E. Stanga, “Electronic Retinal Implants and Artificial Vision: Journey and Present,” *Eye* 31, no. 10 (2017): 1383–1398.
- James D. Weiland and Mark S. Humayun, “Retinal Prosthesis,” *IEEE Transactions on Biomedical Engineering* 61, no. 5 (2014): 1412–1424.
- Eberhart Zrenner, “Will Retinal Implants Restore Vision?” *Science* 295, no. 5557 (2002): 1022–1025.

18See, for example, James J. Jun et al., “Fully Integrated Silicon Probes for High-Density Recording of Neural Activity,” *Nature* 551, no. 7679 (2017): 232–236.

19Paul Bach-y-Rita et al., “Vision Substitution by Tactile Image Projection,” *Nature* 221, no. 5184 (1969): 963–964.

20Hermann von Helmholtz, “The Facts of Perception,” *Selected Writings of Hermann von Helmholtz* (Middletown: Wesleyan University Press, 1971). For a detailed description on this subject, see chapter three of my earlier book *The Forgetting Machine* (Dallas: BenBella Books, 2017).

21Martin Hilbert and Priscilla López, “The World’s Technological Capacity to Store, Communicate, and Compute Information,” *Science* 322, no. 6025 (2011): 60–65.

Chapter 8

1From Jorge Luis Borges, *Ficciones [Fictions] (Buenes Aires: Editorial Sur, 1944).*

2These cases are described in A. Szücs et al., “Misleading Hallucinations in Unrecognized Narcolepsy,” *Acta Psychiatrica Scandinavica* 108, no. 4 (2003): 314–317; and Erin Wamsley et al., “Delusional Confusion of Dreaming and Reality in Narcolepsy,” *Sleep* 37, no. 2 (2014): 419–422.

3Wamsley et al., “Delusional Confusion of Dreaming and Reality in Narcolepsy.

4Translated from *Meditaciones metafísicas con objeciones y respuestas*, sexta Meditación (Madrid: Edición de Vidal Peña, Alfaguara, 1977).

5Thomas Hobbes, “The Third Objections” (response to Descartes’s *Meditations*), 1641/42.

6William H. Moorcroft, *Understanding Sleep and Dreaming* (New York: Springer, 2013), 7.8.

7Miguel Pais-Vieira et al., “A Brain-to-Brain Interface for Real-Time Sharing of Sensorimotor Information,” *Scientific Reports* 3 (2013): 1319.

8Deirdre Barrett, “The ‘Committee of Sleep’: A Study of Dream Incubation for Problem Solving,” *Dreaming* 3, no. 2 (1993): 115–122. See also: Robert Stickgold et al., “Replaying the Game: Hypnagogic Images in Normal and Amnesics,” *Science* 290, no. 5490 (2000): 350–353.

9Sigmund Freud, *Die Traumdeutung* [*The Interpretation of Dreams*] (Leipzig and Vienna: Franz Deuticke, 1900), see chapter seven.

10These are the censorship mechanisms that were missing in Phineas Gage after his frontal lobe injury, creating a radical change in his personality (see chapter one). In line with this evidence, during sleep the activity of the frontal lobe drastically drops. See, for example, Robert Stickgold et al., “Sleep, Learning, and Dreams: Off-line Memory Reprocessing,” *Science* 294, no. 5544 (2001): 1052–1057.

11For a discussion on this subject and the potential errors in psychoanalysis see, for example, Giuliana A. L. Mazzoni and Elizabeth F. Loftus, “When Dreams Become Reality,” *Consciousness and Cognition* 5, no. 27 (1996): 442–462.

12Eugene Aserinsky and Nathaniel Kleitman, “Regularly Occurring Periods of Eye Mobility and Concomitant Phenomena, During Sleep,” *Science* 118, no. 3062 (1953): 273–274.

13Matthew Walker, *Why We Sleep* (London: Penguin, 2018).

14William Dement, “The Effect of Dream Deprivation,” *Science* 131, no. 3415 (1960): 1705–1707.

15J. Allan Hobson and Robert W. McCarley, “The Brain as a Dream State Generator: An Activation-Synthesis Hypothesis of the Dream Process,” *American Journal of Psychiatry* 134, no. 12 (1977): 1335–1348.

16John G. Jenkins and Karl M. Dallenbach, “Obliviscence During Sleep and Waking,” *American Journal of Psychology* 35, no. 4 (1924): 605–612. To review the effects sleeping has on memory, see also:

- Susanne Diekelmann and Jan Born, “The Memory Function of Sleep,” *Nature Reviews Neuroscience* 11, no. 2 (2010): 114–126.
- Matthew Walker and Robert Stickgold, “Sleep-Dependent Learning and Memory Consolidation,” *Neuron* 44, no. 1 (2004): 121–133.

- Stickgold et al., "Sleep, Learning, and Dreams: Off-line Memory Reprocessing."

17Olaf Lahl et al., "An Ultra Short Episode of Sleep Is Sufficient to Promote Declarative Memory Performance," *Journal of Sleep Research* 17, no. 1 (2008): 3–10.

18W. Plihal and J. Born, "Effects of Early and Late Nocturnal Sleep on Declarative and Procedural Memory," *Journal of Cognitive Neuroscience* 9, no. 4 (1997): 534–547.

19To review these results, see: Diekelmann and Born, "The Memory Function of Sleep."

20Stickgold et al., "Sleep, Learning, and Dreams: Off-line Memory Reprocessing."

21Stickgold et al., "Sleep, Learning, and Dreams . . ."

22For a more a elaborate development of this interpretation, see:

- Stickgold et al., "Sleep, Learning, and Dreams . . ."
- Matthew Walker and Robert Stickgold, "Overnight Alchemy: Sleep-Dependent Memory Evolution," *Nature Reviews Neuroscience* 11, no. 3 (2010): 218–219.
- Walker, *Why We Sleep*. In his fascinating book, Walker argues that another function of dreams is processing emotions we have experienced during the day, which is crucial for developing social behaviors (see chapter ten).

Chapter 9

1Plato, *Crito*, 4th century BC.

2Plato, *Theaetetus*, sec. 34.

3Hermann Ebbinghaus, *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* (Leipzig: Duncker and Humblot, 1885).

4James L. McGaugh, "Memory—a Century of Consolidation," *Science* 287, no. 5451 (2000): 248–251.

5James R. Misanin, Ralph R. Miller, and Donald J. Lewis, "Retrograde Amnesia Produced by Electroconvulsive Shock After Reactivation of a Consolidated Memory Trace," *Science* 160, no. 3827 (1968): 554–555.

6Karim Nader, Glenn E. Schafe, and Joseph E. Le Doux, "Fear Memories Require Protein Synthesis in the Amygdala for Reconsolidation After Retrieval," *Nature* 406, no. 6797 (2000): 722–726.

7See, for example, Karim Nader, "Memory Traces Unbound," *Trends in Neurosciences* 26, no. 2 (2003): 65–72; and Karim Nader and Oliver Hardt, "A Single Standard for Memory: The Case for Reconsolidation," *Nature Reviews Neuroscience* 10, no. 3 (2009): 224–234.

8Merel Kindt, Marieke Soeter, and Bram Vervliet, "Beyond Extinction: Erasing Human Fear Responses and Preventing the Return of Fear," *Nature Neuroscience* 12, no. 3 (2009): 256–258.

9To review these results, see Tom Beckers and Merel Kindt, "Memory Reconsolidation Interference as an Emerging Treatment for Emotional Disorders: Strengths, Limitations, Challenges and Opportunities," *Annual Review of Clinical Psychology* 13 (2017): 99–121.

10Elizabeth F. Loftus and John C. Palmer, "Reconstruction of Automobile Destruction: An Example of Interaction Between Language and Memory," *Journal of Verbal Learning and Verbal Behavior* 13, no. 5 (1974): 585–589.

11For more information, see Elizabeth F. Loftus, "Creating False Memories," *Scientific American* 277, no. 3 (1997): 70–75; and Elizabeth F. Loftus and Jacqueline E. Pickrell, "The Formation of False Memories," *Psychiatry Annals* 25, no. 12 (1995): 720–725.

12For more information, see Rodrigo Quian Quiroga, "Magic and Cognitive Neuroscience," *Current Biology* 26, no. 10 (2016): R387–R407 (previously cited in chapter three).

13For more information, see:

- Rodrigo Quian Quiroga, Itzhak Fried, and Christof Koch, "Brain Cells for Grandmother," *Scientific American* 308, no. 2 (2013): 30–35.
- Charles G. Gross, "Genealogy of the 'Grandmother Cell,'" *The Neuroscientist* 8, no. 5 (2002): 512–518.
- Horace Barlow, "The Neuron Doctrine in Perception," *The Cognitive Neurosciences*, edited by Michael S. Gazzaniga (Boston: MIT Press, 1994).
- Rodrigo Quian Quiroga, "Akakhievitch Revisited," *Physics of Life Reviews* 29 (2019): 111–114.

14Jerzy Konorski, *Integrative Activity of the Brain: An Interdisciplinary Approach* (Chicago: University of Chicago Press, 1967).

15Figure adapted from Rodrigo Quian Quiroga, "Neuronal Codes for Visual Perception and Memory," *Neuropsychologia* 83 (2016): 227–241. The neural responses in V1, IT, and the hippocampus were published in:

- David H. Hubel and Torsten N. Wiesel, "Receptive Fields of Single Neurones in the Cat's Striate

Cortex," *The Journal of Physiology* 148, no. 3 (1959): 574–591.

- Charles Gross, "Single Neuron Studies of Inferior Temporal Cortex," *Neuropsychologia* 46, no. 3 (2008): 841–852.
- Rodrigo Quian Quiroga et al., "Explicit Encoding of Multi-Modal Percepts by Single Neurons in the Human Brain," *Current Biology* 19, no. 15 (2009): 1308–1313.

16For a study supporting this claim, see Emanuela de Falco, Matias J. Ison, Itzhak Fried, and Rodrigo Quian Quiroga, "Long-Term Coding of Personal and Universal Associations Underlying the Memory Web in the Human Brain," *Nature Communications* 7 (2016): 13408.

17For more details, see Rodrigo Quian Quiroga, "Concept Cells: The Building Blocks of Declarative Memory Functions," *Nature Reviews Neuroscience* 13, no. 8 (2012): 587–597.

18Matias J. Ison, Rodrigo Quian Quiroga, and Itzhak Fried, "Rapid Encoding of New Memories by Individual Neurons in the Human Brain," *Neuron* 87, no. 1 (2015): 220–230.

19Rodrigo Quian Quiroga, "Concept Cells: The Building Blocks of Declarative Memory Functions."

20For more information, see Karl Deisseroth, "Optogenetics: Controlling the Brain with Light," *Scientific American* 303, no 4 (2010): 48–55; and Karl Deisseroth, "Optogenetics: 10 Years of Microbial Opsins in Neuroscience," *Nature Neuroscience* 18, no. 9 (2015): 1213–1225.

21Edward S. Boyden et al., "Millisecond-Timescale, Genetically Targeted Optical Control of Neural Activity," *Nature Neuroscience* 8, no. 9 (2005): 1263–1268.

22Antoine R. Adamantidis et al., "Neural Substrates of Awakening Probed with Optogenetic Control of Hypocretin Neurons," *Nature* 450, no. 7168 (2007): 420–424.

23Xu Liu et al., "Optogenetic Stimulation of a Hippocampal Engram Activates Fear Memory Recall," *Nature* 484, no. 7394 (2012): 381–385.

24Steve Ramirez et al., "Creating a False Memory in the Hippocampus," *Science* 341, no. 6144 (2013): 387–391.

25To review these results, see Kay M. Tye and Karl Deisseroth, "Optogenetic Investigation of Neural Circuits Underlying Brain Disease in Animal Models," *Nature Reviews Neuroscience* 13, no. 4 (2012): 251–266.

Chapter 10

1See, for example, R. C. Merkle, "The Technical Feasibility of Cryonics," *Medical Hypotheses* 39, no. 1 (1992): 6–16; and Ole Martin Moen, "The Case for Cryonics," *Journal of Medical Ethics* 41, no. 8 (2015): 677–681.

2R. K. Andjus and J. E. Lovelock, "Reanimation of Rats from Body Temperatures between 0 and 1°C by Microwave Diathermy," *The Journal of Physiology* 128, no. 3 (1955): 541–546.

3Robert L. McIntyre and Gregory M. Fahy, "Aldehyde-Stablilized Cryopreservation," *Cryobiology* 71, no. 3 (2015): 448–458.

4*Star Trek*, "The Enemy Within," aired in 1966 and featured the fiendish Kirk replica.

5"Second Chances" from *Star Trek: The Next Generation*, season six, episode twenty-four (aired in 1993).

6Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), chapter ten.

7Parfit, *Reasons and Persons*, 502.

8Daniel C. Dennett, *Consciousness Explained*, p. 418 (cited in chapter two).

INDEX

A

abstraction, 42–43, 233
actions, predicting, 123–124, 263. *see also* free will
agents, 9
Akakhievitch, Akakhi, 217–218, 233
Alcor Life Extension Foundation, 242, 243, 248
alien hand syndrome, 136–137
AlphaGo, 20–22
Amenábar, Alejandro, 240, 241
analogies, 42–43, 233, 263
Andersen, Richard, 156–157
androids (replicants), 25–27, 29, 39. *see also* consciousness
animal consciousness, 47–71
    and awareness of existence, 57
    and coding of concepts in brain, 70–71
    and intelligence in animals, 49–52, 54–59, 265
    and language in animals, 61–69
    philosophers' views of, 48–52, 68
    in *Planet of the Apes*, 47–48, 60–61
    in primates, 54–57, 60–67
    and research on animals, 64–65
    and theory of evolution, 52–54
Aniston, Jennifer, 220–221. *see also* concept cells
An Apology for Raymond Sebond (Montaigne), 50–51
Aquinas, Thomas, 130, 263
Aristotle, 48, 125, 127, 164*n*, 222
artificial intelligence (AI). *see* machine intelligence
Aserinsky, Eugene, 192–193
Asimov, Isaac, 21
associations, 222–225, 232, 234, 265
The Astonishing Hypothesis (Crick), 16
Augustine, Saint, 130
automatic speech recognition, 41
autonomous mind, 150–151, 263
Awake, 108

B

Bach-y-Rita, Paul, 175
Barlow, Horace, 218*n*
Barrett, Deirdre, 189
Bartlett, Fred, 94
Barton, Dave, 225*n*
Baudrillard, Jean, 75*n*
Bennett, Harve, 155
Bergson, Henri, 28, 29

Berkeley, Bishop George, 80, 176
Berry, Halle, 35, 36
binocular rivalry, 30–31
The Bionic Woman, 172–173
Bioy Casares, Adolfo, 84–85
Blade Runner, 25–26, 29, 39, 258
The Book of Dreams (Borges), 182–183
Borges, Jorge Luis, 68, 96, 181–183, 206–207, 265
Boulle, Pierre, 47
brain in a vat argument, 81–82, 85–86, 183
BrainPort, 176, 177
Broad, Charlie Dunbar, 103
Broca, Pierre Paul, 14
Broca’s area, 14–15
butterfly effect, 149

C

Caidin, Martin, 155
Cartesian dualism, 262
    and androids, 29
    and dreams as epiphenomena, 195
    and free will, 151
    and immortality of the mind, 240
    mind–body interaction in, 12–13, 16, 79
causation, 227–230
cell assemblies, 220
Chalmers, David, 27–28
chaos theory, 148–150, 263
chatbots, 40, 41
Cheney, Dorothy, 58
Chinese room argument, 41–42
choice blindness, 151–152
Chomsky, Noam, 61–62, 65
Cicero, 208
“The Circular Ruins” (Borges), 182
Clarke, Arthur C., 7, 9
cochlear implants, 173, 174
COMPAS, 143–145
compatibilism, 130–131
concept cells, 2, 33–35
    and intelligence, 233–235, 265
    and memory, 70, 97, 219–223, 233–235
    reconstructing visual input from, 106–107
consciousness, 25–43. *see also* mind
    and abstractions, 42–43
    in androids, 39 (*see also* androids [replicants])
    in animals (*see* animal consciousness)
    and binocular rivalry, 30–31
    and construction of information, 268–265
    existence of, 262
    experimental paradigms for aspects of, 30
    and fear of death, 258–259
    hippocampal neurons involved in, 35–37
    human, 2
    in machines, 2, 39–42 (*see also* machine consciousness)
    and memory, 36–38
    and physical stimuli giving way to subjective experiences, 31–39
    of self, 257–258

  that precedes action, 137–140 (*see also* free will)
Consciousness Explained (Dennett), 85
consolidation theory, 211–213
Copernicus, Nicolaus, 125, 126
correlation, 227–230
Courtine, Grégoire, 170, 171
creativity, 43, 199–200
Crichton, Michael, 66
Crick, Francis, 16, 30, 227, 228
criminal behavior, 140–147
Crito (Plato), 10
cryonics, 240–248
Cyborg (Caidin), 155
cyborgs, 155–178
  in *The Bionic Woman*, 172–173
  and continuity of the body, 164–166
  and direct muscle stimulation, 172
  ethical issue with, 160
  and exoskeletons, 168–170
  and increasing sensory abilities, 176–177
  and neural prosthetics, 156–162
  in *RoboCop*, 156, 162–166, 168
  robots vs., 156
  and rubber hand illusion, 166–168
  and sensory organ replacements, 172–176
  and Ship of Theseus paradox, 163–164
  in *The Six Million Dollar Man*, 155–156
  and spinal cord electric stimulation, 170–172

D
Dalai Lama, 65
Dallenbach, Karl, 197–198
Darwin, Charles, 52–54, 58–59
De Anima (On the Soul; Aristotle), 48
death, 257–260, 267. *see also* immortality
decoding algorithms, 106–108, 158–159, 161, 162, 170–172
Deep Blue (supercomputer), 17–18
deep neural networks, 18–21, 32
De memoria et reminiscentia (Aristotle), 222
Democritus, 79, 130
Dennett, Daniel, 67*n*
  on Cartesian dualism, 29
  on identity, 253
  on knowledge argument, 104–105
  on reality, 85, 86, 89, 183
  RoboCop reference to, 162*n*
De Oratore (Cicero), 208
Descartes, René, 14
  on advances in science, 127–128
  on animal intelligence/consciousness, 51, 52
  on doubting reality, 183
  on free will, 130–132, 150, 151
  on mind–body/matter division, 2, 12
  philosophy revolutionized by, 261
  on reality, 79, 80
  on sleep vs. wakefulness, 184, 185
The Descent of Man (Darwin), 59
determinism, 128–131, 149

Dick, Philip K., 25, 123
direct muscle stimulation, 172
Discourse on the Method (Descartes), 51, 128
Do Androids Dream of Electric Sheep? (Dick), 25
Dr. Strangelove, 136
dream incubation, 189
dreams, 181–201
  and creativity, 199–200
  dream incubation, 189
  as epiphenomena, 194–196
  Freud's theory of, 190–193
  Hobson's and McCarley's work on, 193–194
  in *Inception*, 181–184, 186–188, 200–201
  Jung's view of, 193
  lucid, 186–189
  memory and construction of, 198–199
  in *One Hundred Years of Solitude*, 196–197
  as premonitions, 189–190
  and preserved brains, 244
  reality vs., 182–183
  and REM sleep, 192–193, 198–199
  and role of sleep in memory, 197–198
  sharing, 188
  wakefulness vs., 183–186
Dreyfus, Hubert, 162*n*

E
Eagleman, David, 177*n*
Ebbinghaus, Hermann, 210–211
Ehrsson, Henrik, 167
Ekman, Paul, 118*n*
El Tamaño De Mi Esperanza (Borges), 68
emotions, 56–57, 65, 117–118
The Emperor's New Mind (Penrose), 129–130*n*
encoding memories, 210
engrams, 220, 234
epilepsy, 15, 32, 33, 82–83
epiphenomena, dreams as, 194–196
episodic memory, 226*n*, 234
Essai philosophique sur les probabilités (Laplace), 128
An Essay Concerning Human Understanding (Locke), 68, 176
Essay Towards a New Theory of Vision (Berkeley), 176
Eternal Sunshine of the Spotless Mind, 225
Ethics (Spinoza), 132–133
Ettinger, Robert, 242*n*
evil genius argument, 80, 85, 128
evolution, 52–54
exoskeletons, 168–170
explanatory gap (the hard problem), 27–28, 30, 195–196, 256, 262
The Expression of the Emotions in Man and Animals (Darwin), 54

F
Facebook, 19
face patches, 116–117, 233*n*
"The Facts of Perception" (Helmholtz), 90
Fallon, Jim, 140–141
Fassbinder, R. W., 101

fear conditioning, 212–214
flashbulb memory, 77
free will, 123–152, 263
  and alien hand syndrome, 136–137
  and autonomous mind, 150–151
  and chaos theory, 148–150
  and choice blindness, 151–152
  and consciousness that precedes action, 137–140
  and criminal behavior, 140–147
  and determinism, 128–131
  and electrical stimulation of brain neurons, 134–135
  illusion of, 135–138
  and Libet's readiness potential, 131–133
  in *Minority Report*, 123–124, 141, 145, 146
  and moral decisions, 140
  philosophical debates about, 130–133
  and scientific revolution, 125–128
  and voxel activation patterns, 133–134
Freiwald, Winrich, 116–117, 219
Freud, Sigmund, 190–194, 218*n*
Fried, Itzhak, 84–85, 134
Front of the Class, 139*n*
functionalism, 249
"Funes the Memorious" (Borges), 96

G
Gabriel, Peter, 66, 101
Gage, Phineas, 14, 15
Galilei, Galileo, 125, 127
Gallant, Jack, 113–114
Gallup, Gordon, Jr., 57
García Márquez, Gabriel, 196–197
Gardner, Allen, 61, 62
Gardner, Beatrix, 61, 62
Gatti, Hugo, 95–96
Gazzaniga, Michael, 141
Gea, Miguel Ángel, 92, 217
Geller, Uri, 13*n*
general intelligence, 22, 43, 71, 233, 264, 268
generalizing, 42–43
Genesis, Book of, 189–190
Gibson, William, 77
"Gilgamesh, the Immortal," 239, 259–260
Goodall, Jane, 57–58, *64n*
Google, 19, 20
Gross, Charles, 219

H
HAL (Hybrid Assistive Limb) exoskeleton, 169
Handke, Peter, 101–102
Hasson, Uri, 111–112
Hauser, Kaspar, 69
Haxby, Jim, 112–113
Hayes, Catherine, 61
Hayes, Keith, 61
Haynes, John-Dylan, 133
Helmholtz, Hermann von, 89–90, 176, 265

Her, 40*n*
Heraclitus, 164*n*
Herzog, Werner, 101
Hinton, Geoffrey, 18–19
Hitchcock, Alfred, 91
Hobbes, Thomas, 185
Hobson, Allan, 193–194, 199
Houdini, Harry, 92–94
Howe, Elias, 200
Hubel, David, 219
human(s)
    and androids, 25–26
    consciousness of, 2
    and machines, 12, 264
Human Knowledge (Russell), 81
Humans (TV series), 27
Hume, David, 94, 253, 266
hypothermia, 246–248

I

I, Robot (Asimov), 21
IBM, 19
identity, 178. *see also* self
    after cryopreservation, 246–248
    after reproducing the brain, 249
    after teleportation, 250–252
    as construction of brain, 257, 265–266
    and continuity of the body, 164–166
    and fear of death, 257–259
    as illusion, 253–256
    and memory, 165–166, 178, 206–207, 252, 266
    and organ transplants, 178
    and rubber hand illusion, 166–168
    and Ship of Theseus paradox, 163–164
illusion of reality, 75–97
    and brain function, 96
    created by magicians, 92–94
    and existence of others, 81, 86
    and language, 87–89
    in *The Matrix*, 75–78, 80–81, 89, 92, 97
    and perceptions altered by brain stimulation, 84
    personal interactions within, 84–86
    philosophers' views on, 78–90
    and schemas, 94–97
    and unconscious inferences, 89–92, 94–97
imitation game, 39–40
immortality, 239–260, 266–267
    and cryopreservation, 240–246
    and fear of death, 257–259
    and feeling of our own existence, 255–256
    in "Gilgamesh, the Immortal," 239
    and identity, 246–248, 250–252, 256–257
    of the mind, 240
    in *Open Your Eyes*, 240–242
    by reproducing the brain, 249
    and self as illusion, 252–255
    and survival instinct, 239
Inception, 1, 181–184, 186–188, 200–201

inferences, 233, 263. *see also* unconscious inferences
Institutio Oratoria (Quintilian), 208
intelligence
    in animals, 49–52, 54–59, 69–70, 265 (*see also* animal consciousness)
    and assumption-making, 92–94
    and concept cells, 233–235
    cultural influences on, 59–60
    general, 22, 43, 71, 233, 264, 268
    in humans vs. in machines, 9, 10, 17, 22
    and language in animals, 51–52, 61–69
    of machines (*see* machine [artificial] intelligence)
The Invention of Morel (Bioy Casares), 84–85
Irvine, Andrew, 247, 248

J
Jackson, Frank, 103, 104
James, William, 218, 222
Jenkins, John, 197–198
“Jennifer Aniston neurons.” *see* concept cells
Jolie, Angelina, 35, 36
Jung, Carl, 191, 193, 194

K
Kafka, Franz, 165
Kamitani, Yukiyasu, 115, 188
Kant, Immanuel, 80, 140
Kasparov, Garry, 17–18
Kekulé, August, 200
Kepler, Johannes, 127
Kindt, Merel, 214
Kleitman, Nathaniel, 192
knowledge argument, 103–105
Koch, Christof, 16*n*, 30, 156–157
Köhler, Wolfgang, 54–55
Konorski, Jerzy, 219
Kubrick, Stanley, 7, 9, 136
Kudrow, Lisa, 221

L
LaBerge, Stephen, 187
LaFarge, Stephanie, 63, 64
language, 15*n*, 265
    abstraction in, 67–69
    in animals, 51–52, 61–69
    and coding of concepts in brain, 70–71
    and existence of others, 88–89
    and humans’ need to express thoughts, 67
    and meaning of words, 87–88
The Language Instinct (Pinker), 62
Laplace, Pierre-Simon, 128–129
La planète des singes (Boulle), 47
Lavand, René, 93*n*
Leborgne, Louis Victor, 14, 15
Leibniz, Gottfried, 28, 29, 262–263
Lettvin, Jerry, 218
Libet, Benjamin, 131–133
“Lied vom Kindsein” (Handke), 101–102

life, criteria for, 28–29
The Little Prince (Saint-Exupéry), 64
Locke, John, 68, 165, 176, 207
locked-in syndrome, 107–108
Loebner Prize, 40
Loftus, Elizabeth, 214–216
Logothetis, Nikos, 31
long-term memory, 210–211
Loomis, Eric, 144
Lorenz, Edward, 147–149
"The Lost Mariner" (Sacks), 38*n*
lucid dreams, 186–189

M

machine consciousness, 2, 39–42, 258, 267–268. *see also* androids (replicants)
machine (artificial) intelligence, 7–22, 263–265
    advances in, 17–21
    and Chinese room argument, 41–42
    consequences of advances in, 21–22
    cyborg abilities vs., 156
    deep neural networks, 18–21
    and human intelligence, 9, 10, 17, 22
    and materialism, 17
    and mind–body/matter question, 10–16
    Minsky on, 9–10
    paradigm shift in, 18–19
    SNARC artificial neural network, 9
    in *2001: A Space Odyssey*, 7–9
magic and magicians, 57*n*, 92–94, 216–217
magic realism, 196–197
Maldonado, Héctor, 208*n*
Mallory, George, 247, 248
materialism, 262
    and dreams, 195–196
    and knowledge argument, 103–105
    and machine intelligence, 17
    and mind–body connection, 240, 262
    modern view of, 79–80
    and replication of mind, 30
The Matrix, 75–78, 80–81, 85, 92, 97, 183, 242
Maxwell, James Clerk, 13
McCarley, Robert, 193–194, 199
McCartney, Paul, 200
Meditations on First Philosophy (Descartes), 12, 79, 184, 185
Memento, 37, 38
memory(-ies), 205–235
    activated by brain stimulation, 83
    after reproducing the brain, 249, 252
    after teleportation, 252
    in ancient times, 208–210
    in androids, 26
    associations in, 222–225, 232
    brought to consciousness, 36–38
    and concept cells, 70, 219–223, 233–235, 265
    consolidation of, 211–213
    as construction of brain, 265–266
    and correlation vs. causation, 227–230
    creating, 230

  and dream construction, 198–199
  erasing, 211, 213–214, 217–218, 225–227
  false, 214–216, 232
  first systematic study of, 210–211
  flashbulb, 77
  in H.M. case, 15, 32
  and identity, 165–166, 178, 206–207, 252, 256, 266
  implanted, 205–207, 211, 215–216, 230, 232–235
  of jays, 56
  long- and short-term, 210–211
  manipulating, 211, 213–217, 222–224, 229–232
  and optogenetics, 229–232
  reconsolidation of, 213–216
  role of sleep in, 197–198
  schemas or inferences in, 94–97
  storage of, 222–225
  in *Total Recall*, 205–206, 232
  unconscious inferences in, 224–225
  understanding vs., 43
  and visual processing, 219–222
Metamorphosis (Kafka), 165
Metropolis, 250*n*
Microsoft, 19
Milner, Brenda, 37
mind, 2, 10. *see also* consciousness
  as activity of neurons, 107
  autonomous, 150–151, 263
  brain injuries that change, 14–16
  immortality of, 240 (*see also* immortality)
  theory of, 56–57
mind–body/matter question, 10–16, 262. *see also specific topics; specific topics, e.g.:* immortality
mind reading. *see* reading the mind
Minority Report, 123–124, 141, 145, 146
Minsky, Marvin, 9–10, 17, 21–22, 40, 42
mixed aphasia, 15*n*
Mnemosyne, 208
Molaison, Henry (H.M.), 15, 32, 36–38, 225–226
Molina, Mariano, 31*n*
Monadology (Leibniz), 28
monism, 79–80
Montaigne, Michel de, 50–53, 55
Moon, 255, 256
Moores, Ezequiel Fernández, 95
Moralia (Matters Relating to Customs and Mores; Plutarch), 49
motor aphasia, 15*n*
movies, 1
Müller, Georg, 211

N
Nagel, Thomas, 104
narcolepsy, 184, 185
neural networks, 9, 18–21
neural prosthetics, 156–162
Neuromancer (Gibson), 77
neuroplasticity, 161–162
neuroscience revolution, 1–2
Newton, Isaac, 43, 125, 127, 200
Nicolelis, Miguel, 168, 172, 188

Nolan, Christopher, 181, 183, 186, 201
NREM (non-rapid eye movement) sleep, 192–193, 198–199

O
Oblivion, 255, 256
obstetric dilemma, 71*n*
Olivera, Lucho, 239, 260
One Hundred Years of Solitude (García Márquez), 196–197
On the Origin of Species (Darwin), 52
ontological idealism, 79, 80
Open Your Eyes, 240–242
optogenetics, 229–232
organ transplants, 178
Outlines of Pyrrhonism (Sextus Empiricus), 49–50, 67
Owen, Adrian, 109–112

P
pain, 27, 262–263
paper manufacturing, 209
paradoxical sleep, 192
Parfit, Derek, 250–253, 256
Parmenides, 164*n*
Pascual-Leone, Álvaro, 134
Penfield, Wilder, 82–83, 134, 135
Penrose, Roger, 129–130*n*
Pepperberg, Irene, 52*n*
perception(s), 28, 265
  altered by brain stimulation, 82–84
  of change, 164*n*
  and existence, 80
  in movies about angels, 102–103
  of reality, 261
  and response of world to expectations, 85–86
  and unconscious inferences, 89–92, 94–97
Perry, John, 255
Phaedo (Plato), 10–12, 164*n*
Philosophiæ Naturalis Principia Mathematica (Newton), 125
Philosophical Investigations (Wittgenstein), 87, 88
philosophical zombies (p-zombies), 26–27, 38
philosophy, 2
Pilzecker, Alfons, 211
Pinker, Steven, 62
place cells, 70
Planet of the Apes, 47–48, 60–61
Plato, 10–12, 78–77, 164*n*, 208–210
platonic realism, 79
Plutarch, 49–52, 55, 163, 164
Poincaré, Henri, 148*n*
pontifical cells, 218–219
Portal, Raúl, 13*n*
post-traumatic stress disorder, 213, 214
private language, 88
Project Nim (documentary), 63, 64
proof of principle, 159
The Prospect of Immortality (Ettinger), 242*n*
Ptolemy, 125
Putnam, Hilary, 81–82

Q
qualia. *see* subjective experiences
quantum mechanics, 129–130*n*
Quintilian, 208

R
Ramón y Cajal, Santiago, 191–192, 218*n*
Randi, James, 13*n*
Rashid, Rick, 41
rational soul, 48
readiness potential, 131–133
reading the mind, 101–119
  decoding algorithms in, 106–108
  and determination of consciousness, 108–111
  by identifying brain activation areas, 111–112
  by interpreting emotions from facial expressions, 117–118
  by interpreting unconscious gestures, 118, 119
  and knowledge argument, 103–105
  to reconstruct thoughts or dreams, 114–115, 117
  and research on face patches, 116–117
  in *Until the End of the World*, 101, 105–106, 115
  and vegetative states, 108–111
  from voxel activation pattern in visual areas, 112–115
  in *Wings of Desire*, 101–105
reality, 182–183, 261. *see also* illusion of reality
recalling memories, 210, 211, 213
Reeve, Christopher, 158
remapping, 70
REM (rapid eye movement) sleep, 192–193, 198–199
reproducing the brain, 249
The Republic: Book VII (Plato), 78–77
retinal implants, 173–175
retributive justice, 142, 143
Rhetorica ad Herennium, 208
Rise of the Planet of the Apes, 63
RoboCop, 156, 162–166, 168
rubber hand illusion, 166–168
Russell, Bertrand, 81
Ryle, Gilbert, 16

S
Sacks, Oliver, 38*n*
Saint-Exupéry, Antoine de, 64
Savage-Rumbaugh, Sue, 65–66
schemas, 94–97
Schiavo, Terri, 108–109
schizophrenia, 137, 139
Schwartz, Andy, 159
science fiction, 2
scientific revolution, 125–128
sci-fi movies, 1
Scott, Ridley, 25, 26*n*
Scoville, William, 15
Searle, John, 41–42
"The Secret Miracle" (Borges), 181–182
Sedol, Lee, 20–22
self, 252–259, 266–268. *see also* identity

self-awareness, 57, 255. *see also* consciousness
semantic memory, 226*n*
sensitive soul, 48
sensory aphasia, 15*n*
sensory organs, 172–177
The Seventh Seal, 240, 241
Sextus Empiricus, 49–50, 67
Seyfard, Robert, 58
“Shakespeare’s Memory” (Borges), 206–207
Shelley, Mary, 200
Sherrington, Charles, 218
Ship of Theseus paradox, 163–164, 248–249
short-term memory, 210–211
Simulacra and Simulation (Baudrillard), 75*n*
The Six Million Dollar Man, 155–156
sleep
    and cryopreservation, 244
    eye movement during, 192–193
    and memory, 197–198
    stages of, 184, 185, 192–193, 198–199
    testing reality during, 186
    time spent in, 189
SNARC artificial neural network, 9
The Society of Mind (Minsky), 9
Socrates, 10–11
solipsism, 80–81, 86, 89
sorites paradoxes, 164
Spielberg, Steven, 123, 124
spinal cord electric stimulation, 170–172
Spinoza, Baruch, 132–133, 267
Star Trek, 250
Star Wars, 254, 255
Stevenson, Robert Louis, 200
Stickgold, Robert, 198–199
Stoic philosophers, 48–49
subjective experiences (qualia), 27–28
    and feeling of existing, 256
    and knowledge argument, 104–105
    physical stimuli giving way to, 31–39, 196
survival instinct, 239

T

teleportation, 250–252, 256–257
Terminator 2, 21, 27, 258
Terrace, Herbert, 62, 63, 65
theory of mind, 56–57
thinking. *see also* consciousness
    ability to generalize and use analogies, 42–43
    by animals (*see* animal consciousness)
    by machines, 10, 41–42 (*see also* androids [replicants]; machine [artificial] intelligence)
thoughts, 3
    interaction of matter and, 13*n*
    need to express, 67
    of others, understanding, 56–57
    predicting, 107 (*see also* reading the mind)
    reconstructing, 114–115, 117
“Throat of a Black Kitten,” 135–136
“Tlön, Uqbar, Orbis Tertius” (Borges), 68

Tonegawa, Susumu, 230–232
Total Recall, 205–206, 232, 252
Tourette syndrome, 139
Tractatus Logico-Philosophicus (Wittgenstein), 87
transcendental idealism, 80
transcranial magnetic stimulation (TMS), 134
A Treatise of Human Nature (Hume), 253
Tsao, Doris, 116–117, 219
Turing, Alan, 39–40, 42, 263
Turing test, 39–41, 263, 267
2001: A Space Odyssey, 1, 7–9, 17, 47

U
unconscious inferences, 89–92, 94–97, 224–225, 265
Until the End of the World, 1, 101, 105–106, 115
utilitarian justice, 142–143

V
Vanilla Sky, 241*n*
vegetative soul, 48
vegetative states, 108–111
Vernon, Dai, 92–94
visual illusions, 90–91
visual processing in brain, 32–37, 219–222
vOICe, 176, 177

W
Wachowski siblings, 75*n*
Watson, James, 16*n*
Watson supercomputer, 41, 42
Wenders, Wim, 101, 102, 105, 115
Wernicke, Carl, 15*n*
Wernicke's area, 15*n*
Wiesel, Torsten, 219
Wilkins, Maurice, 16*n*
Williams, Bernard, 165*n*
Wings of Desire, 101–105
Wittgenstein, Ludwig, 67, 86–89, 183

Z
Zeno, 182*n*
Zhuang Zhou, 182–183
Zilly, Paul, 144

ABOUT THE AUTHOR

Photo courtesy of Univeristy of Leicester.

RODRIGO QUIAN QUIROGA holds a Research Chair at the University of Leicester, where he is the director of the Centre for Systems Neuroscience. He graduated in Physics at the University of Buenos Aires, Argentina, and obtained his PhD in Applied Mathematics at the University of Luebeck, Germany. Before joining the University of Leicester in 2004, he was a post-doctoral fellow at Germany's Research Center Juelich, a Sloan fellow at the California Institute of Technology, and had short stays at RIKEN in Japan and the University of Nijmegen in The Netherlands. He has held visiting positions at the Leibniz Institute for Neurobiology in Magdeburg; the International School for Advanced Studies (SISSA) in Trieste, Italy; the University of California, Los Angeles; the California Institute of Technology; Fundacion Pascual Maragall in Barcelona; Peng Cheng Lab in China; the University of Buenos Aires; and CONICET in Argentina. He is the recipient of a young investigator award from the American Epilepsy Society and a Royal Society Wolfson Research Merit Award. In 2014 he was selected as one of the 10 UK RISE Leaders in Science and Engineering, and since 2019 has been a member of the Academy of Medical Sciences.

His primary research interest is in visual perception, memory, and the bases of human thought and intelligence. He discovered what have been named "concept cells" or "Jennifer Aniston neurons"—neurons in the human brain that play a key role in memory—a finding that was selected as one of the top 100 scientific stories of 2005 by *Discover Magazine*. His work has been published in over 100 research articles and has received worldwide media attention, including in the *New York Times, Washington Post, Scientific*

American, New Scientist, and *The Independent*, among others. He is the author of four other books, including *Borges and Memory*, linking the ideas of Argentine writer Jorge Luis Borges with the neuroscience of memory.